Statistical Analysis of Multivariate Infectious Disease Surveillance Time Series

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Michaela Paul

aus

Deutschland

Promotionskomitee

Prof. Dr. Leonhard Held (Vorsitz)
Prof. Dr. Andrew Barbour
Prof. Dr. Peter Bühlmann (ETH Zürich)

Zürich, 2011

Preface

I am grateful to many people for all the support and encouragement I have received during the last years.

First of all I would like to thank my supervisor Leonhard Held for many valuable discussions and suggestions whilst working on my thesis. I am very grateful for his support and encouragement to participate in national and international conferences and to spend a research visit in the UK. My thanks also go to my co-authors, in particular Michael Höhle and Sereina Herzog, for all the constructive comments and the time they spent discussing with me. I owe thanks to all my colleagues at the Biostatistics Unit for their support, be it discussions and comments, proofreading of parts of this thesis, administrative help or encouragement. Especially, I would like to thank Andrea Riebler for sharing office with me and providing discussion and help whenever needed. I am also grateful to Paddy Farrington for his hospitality and many useful discussions during my stay at the Open University in Milton Keynes. Furthermore, I would like to thank Peter Diggle for agreeing to review my thesis, and Andrew Barbour and Peter Bühlmann for being part of my dissertation committee. Financial support by the Swiss National Science Foundation is gratefully acknowledged.

Finally, I would like to thank my family for their support in any respect throughout my dissertation.

Zurich, October 2010

Michaela Paul

Zusammenfassung

Zum wirkungsvollen Schutz der Gesundheit der Bevölkerung haben viele Länder nationale Überwachungssysteme aufgebaut, die fortlaufend Daten für eine Reihe meldepflichtiger Krankheiten sammeln. Die Analyse und Modellierung solcher Meldedaten trägt wesentlich dazu bei, die Ausbreitung von Krankheiten zu begrenzen und zu verhindern. In der vorliegenden Arbeit wird statistische Methodik für die Analyse von multivariaten Zeitreihen von Zähldaten entwickelt, wie sie von Überwachungseinheiten im Rahmen des Infektionsschutzes gesammelt werden. Ziel dieser Dissertation ist es ein flexibles Modell bereitzustellen mit dem sowohl zeitliche und räumlich-zeitliche Trends in den Daten, als auch Abhängigkeiten zwischen verschiedenen Krankheiten erklärt werden können. Um eine einfache Benutzung der Modelle zu gewährleisten, wurde sämtliche Methodik im frei erhältlichen R-Paket 'surveillance' implementiert.

Die Modellierung von Ausbrüchen und Unregelmäßigkeiten in den Daten stellt eine besondere Herausforderung dar. Ausgehend von einem Verzweigungsprozess wird dies mittels einer autoregressiven Formulierung erreicht. Mögliche Abhängigkeiten zwischen mehreren Krankheiten können durch die Einführung krankheitsspezifischer Parameter untersucht werden. Die Analyse von wöchentlichen Influenza und Meningokokken Fallzahlen weist empirisch darauf hin, dass eine vorangehende Influenza-Erkrankung die Infektion mit Meningokokken begünstigt. Um die räumliche Ausbreitung einer Krankheit besser widerzuspiegeln, können in einem räumlich-zeitlichen Kontext auch externe Daten, wie z.B. Informationen über das Reiseverhalten, genutzt werden.

Die Formulierung des nicht-linearen autoregressiven Modells wird erweitert um regionale Unterschiede bezüglich der Übertragung und Inzidenz bei in hohem Maße stratifizierten Zeitreihen berücksichtigen zu können. Regionale Heterogenität kann durch Unterschiede in der Alters- und Geschlechtsstruktur, dem Impfstatus der Bevölkerung oder der Umweltbedingungen in der jeweiligen Region verursacht werden. Abhängig davon ob geeignete Information bezüglich solcher Faktoren vorhanden ist oder nicht werden zwei Modellerweiterungen entwickelt.

Der erste Ansatz schließt unkorrelierte und räumlich korrelierte zufällige Effekte im Modell mit ein. Inferenz für dieses Modell basiert auf einem penalisierten Likelihood Ansatz. Varianz Parameter werden mittels marginaler Likelihood geschätzt. Für die Schätzung korrelierter zufälliger Effekte wird das Schätzverfahren entsprechend angepasst. Da klassische Modellwahlkriterien wie AIC oder BIC bei Vorliegen zufälliger Effekte problematisch sein können, basiert die Modellwahl auf Ein-Schritt-Vorhersagen und Proper Scoring Rules. In zwei Anwendungen wird gezeigt, dass sich die prädiktive Güte bei Berücksichtigung von bestehender Heterogenität mittels zufälliger Effekte verbessert.

Wenn Heterogenität durch bekannte und beobachtbare Faktoren bedingt ist kann diese Information mit Hilfe eines Regressionsansatzes direkt in Bezug zu den autoregressiven Parametern gesetzt werden. Solch ein Regressionsansatz ermöglicht auch die Spezifikation zeitlich variierender autoregressive Parameter, beispielsweise um die Wirkung von Interventionen abzubilden. In einer Anwendung auf Masern Surveillance Daten aus Deutschland werden länderspezifische Durchimpfungsraten genutzt um regionale Unterschiede im Verlauf der Inzidenz zu erklären.

Abstract

To meet the threats of infectious diseases, many countries have established surveillance systems for the routine collection of infectious disease data. The analysis and modelling such notification data is essential in the attempt to control and prevent disease. In this thesis, statistical methodology for the analysis of multivariate time series of counts as collected in surveillance systems on notifiable diseases is developed. The aim is to provide a flexible model which is able to explain the temporal and spatio-temporal patterns in the data, as well as account for dependencies between different pathogens. The methodology is implemented in the open-source R-package 'surveillance' which facilitates its use in practice.

A particular challenge is the modelling of outbreaks and irregularities in the data. Motivated by a branching process formulation, well-known in infectious disease epidemiology, the epidemic behavior is modelled via an autoregressive formulation. Possible dependencies between related diseases are analyzed by introducing disease-specific parameters. An analysis of weekly influenza and meningococcal disease counts shows empirical evidence that influenza infections predispose meningococcal disease. In a spatio-temporal context, we propose incorporating external data such as travel intensities between regions to better reflect the regional spread of a disease.

The non-linear autoregressive model formulation is further extended to integrate regional heterogeneity in disease transmission and incidence for highly multivariate time series. Such differences may be due to age, sex, vaccination status or environmental conditions. Two approaches to address heterogeneity are developed, depending on whether or not suitable covariate information about such factors is available.

In the first approach, uncorrelated and spatially correlated random effects are included in the model. The random effects may describe heterogeneity in disease incidence levels as well as in the autoregressive coefficients, relating disease incidence to past counts in the same or neighboring regions. Inference for this non-standard model is based on penalized likelihood methodology. Variance parameters are estimated using marginal likelihood. The estimation procedure is adapted to handle correlated random effects. As classical model choice criteria such as AIC or BIC can be problematic in the presence of random effects, model choice is performed using one-step-ahead predictions and proper scoring rules. As exemplified by two applications, the predictive performance improves if existing heterogeneity is accounted for via random effects.

When heterogeneity is due to known and observable factors, this information can be directly related to for example the autoregressive parameters via a regression formulation. Such a formulation also permits the autoregressive parameter to vary over time, for instance to reflect public health interventions. In an application to German measles surveillance data, region-specific vaccination coverage levels are used to explain regional differences in the incidence pattern.

Thesis outline

Introduction

Paper I: **Statistical approaches to the monitoring and surveillance of infectious diseases for veterinary public health**

Michael Höhle, Michaela Paul & Leonhard Held

Paper published in *Preventive Veterinary Medicine* 2009, **91**, 2–10.

Paper II: **Multivariate modelling of infectious disease surveillance data**

Michaela Paul, Leonhard Held & André M. Toschke

Paper published in *Statistics in Medicine*, 2008, **27**, 6250–6267.

Paper III: **Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts**

Michaela Paul & Leonhard Held

Paper accepted for publication in *Statistics in Medicine*.

Paper IV: **Heterogeneity in vaccination coverage explains the size and occurrence of measles epidemics in German surveillance data**

Sereina A. Herzog, Michaela Paul & Leonhard Held

Paper published in *Epidemiology and Infection*, 2010 (in press).

Appendix I: **Incorporating random effects in a likelihood-based model for multivariate infectious disease counts**

Michaela Paul & Leonhard Held

Extended abstract published in the Proceedings of the 16^{th} *European Young Statisticians meeting*, Bucharest, Romania, 2009.

Appendix II: **Inference details**

Appendix III: **Software manual**

Introduction

Infectious diseases pose a major threat to the health of humans and animals. The availability of vaccines and antibiotics has lessened the disease burden, while newly emerging pathogens such as HIV or SARS, as well as existing pathogens such as malaria, continue to cause major morbidity and mortality (M'ikanatha, Lynfield, Julian, Van Beneden and de Valk, 2007). Since the end of the 19th century, many countries have thus established surveillance systems for the reporting of various infectious diseases to control and prevent the spread of infections (Giesecke, 2002, Chapter 13). The systematic and standardized reporting at a national and regional level aims to recognize all outbreaks quickly, even when aberrant cases are dispersed in space.

Traditionally, notification data, i.e. counts of cases confirmed according to a specific definition and reported daily, weekly or monthly on a regional or national level, are used for surveillance purposes. Other data sources include hospital admission data or syndromic data such as over-the-counter sales of drugs. For illustration, Figure 1 shows weekly notification data for four selected diseases reported to the Robert Koch Institute (RKI) as part of the German 'Protection Against Infection Act' (Infektionsschutzgesetz, IfSG). Each time series comprises the total number of laboratory confirmed cases of a specific disease in Germany, and may be further stratified by e.g. region or age. As seen from the figure, many diseases undergo considerable seasonal variation, while others also display long-term trends. A typical feature is the occasional occurrence of irregularities (outbreaks).

The modelling of infectious diseases is central to our understanding of pathogen evolution and ecology as well as to the prediction of disease dynamics, and is a precondition for effective surveillance. In this thesis, a statistical modelling and inference approach for the analysis of multiple time series of reported counts, as exemplified in Figure 1, is developed. The primary objective is to identify outbreaks and spatio-temporal patterns, and to explore interdependencies between different pathogens using a flexible statistical model. The proposed count data model accounts for serial and spatio-temporal correlation, as well as for heterogeneity in incidence levels and disease transmission. Statistical inference is based on (penalized) likelihood methodology, software for estimating and comparing the models was implemented in the freely available `R` (R Development Core Team, 2010) package `surveillance` (Höhle, 2007). Particular focus is on the analysis of person-to-person transmitted communicable infections such as influenza or common childhood infections. This is mainly done in a retrospective manner to explore past outbreaks and diseases patterns for a fixed data set. Furthermore, the model can be used as a predictive tool and eventually may be applied to prospective (on-line) detection of outbreaks. Here, data are recorded sequentially over time, and the decision concerning whether or not incidence has increased is made sequentially, based on the data collected thus far.

In the following, an overview of various mechanistic and empirical models that are used to study infectious diseases is given. Mechanistic models are based on our understanding of the underlying epidemiological processes and aim to describe the disease transmission mechanism, while empirical models are mainly specified by exploring data and aim to explain the variability in the observations (Pawitan, 2001, Section 1.2). Of course, the distinction between those two classes is not clear-cut and there is a range of intermediate models which contain both mechanistic and empirical components. The use of different models as tools for early outbreak detection is outlined. Finally, the count time series based modelling approach, which forms the basis for this thesis, is introduced.

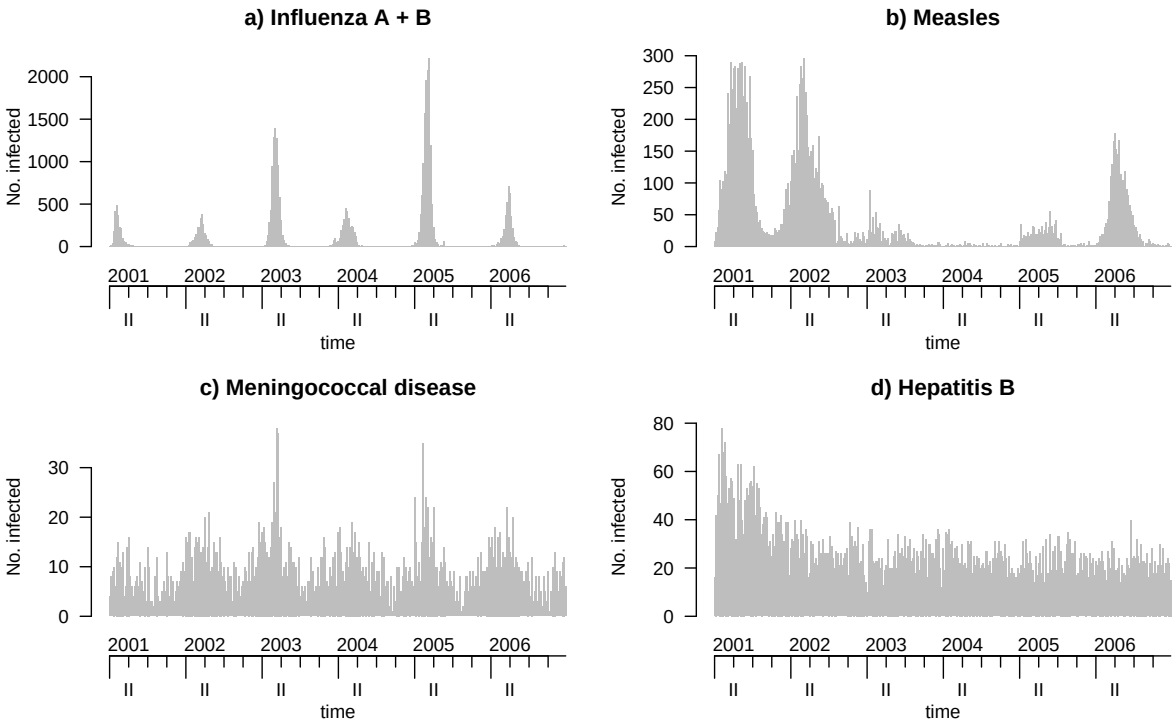


Figure 1.: Weekly number of cases of four diseases reported to the Robert Koch Institute, Germany, in the years 2001–2006 as part of the IfSG. Data were obtained from `http://www3.rki.de/SurvStat`.

1 Analysis of infectious diseases

There is one unique feature of infectious diseases that makes the modelling different from modelling other non-infectious diseases (see the introductory chapters by Giesecke, 2002; or Becker, 1989): the infectious disease can be acquired by contact with an infectious agent. Transmission may either occur directly between individuals or indirectly through the environment (e.g. contaminated water or food) or intermediate hosts (e.g. mosquitos). The majority of viral and bacterial infections—such as measles, chickenpox, HIV, hepatitis B, influenza or meningococcal disease—are directly transmitted.

Figure 2 shows a schematic representation of the disease progress in an individual for an infectious disease which results in lifelong immunity. An example are childhood infections. At some point in time, the susceptible individual comes into contact with the infectious agent. This is the start of the *latent period*, during which the individual is not yet infectious but carries the pathogen. It is followed by the *infectious period*, during which the person can transmit the disease. Symptoms of the disease will appear after the *incubation period*. Note that the infectious and symptomatic periods do not necessarily coincide. Finally, the individual recovers and, in this example, is immune to further infection. For other infections, the individual may be susceptible again after recovery, either immediately or after some time period.

The disease transmission between individuals depends on three factors: the presence of an infectious individual, the presence of a susceptible individual, and the occurrence of effective contacts between those two groups. In short, an epidemic can be described by listing who infected who and when (Grassly and Fraser, 2008). An important quantity for the spread of

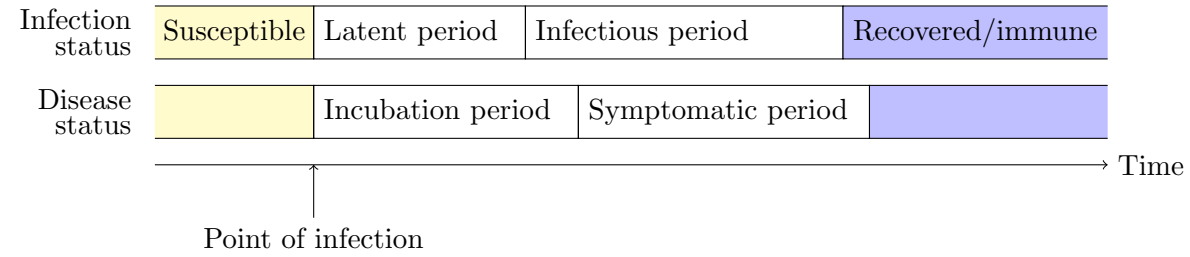


Figure 2.: Schematic illustration of the timeline of infection and disease status in an individual for a hypothetical infection. Here, it is assumed that once an individual is recovered, it is not susceptible anymore. The sum of average latent and average infectious period is called the average generation time of infection (Anderson and May, 1991, p. 14).

the disease in a population is the *average generation time*, that is the average time between successive 'generations' of infectives. It is defined by Anderson and May (1991, p. 14) as the sum of average latent and average infectious period (see also Figure 2).

The epidemic process is usually not completely observable. For instance, it is virtually impossible to know the exact time point at which the infection occurred or how long the infectious period lasted. This makes it very difficult to obtain detailed and precise data, which in turn complicates inference about key parameters even in simple mechanistic models (Andersson and Britton, 2000). Moreover, notification data are typically not individual-based, but aggregated according to a specific partition of gender, age and location. As a consequence of the disease transmission mechanism, the observations in the resulting time series of infectious disease counts are inherently dependent. The analysis of surveillance data may be further complicated by underreporting or other reporting artefacts. Finally, when faced with outbreaks of infectious diseases, there may only be little time for investigations before one has to decide on preventive actions. In this regard, it is particularly advantageous if such decisions can be based on statistical analyses where results are readily available. Complex models requiring e.g. computer-intensive Markov chain Monte Carlo (MCMC) methods may be impractical.

1.1 Mechanistic epidemic modelling

Compartmental models

Depending on an individual's ability to transmit the infectious agent, the population may be divided into three compartments of individuals (see Figure 2): those susceptible to infection (S), those currently infectious (I), and those recovered (R) and immune to infection. This fundamental classification into a few compartments builds the basis for most epidemic models. Greater detail and realism can be achieved by adding additional compartments such as e.g. a class of exposed (E), i.e. latent but not yet infectious individuals. For other infections, where it is impossible to acquire immunity, the class R of immune individuals may be omitted. A full model for the dynamics of the disease is obtained by specifying transition rates, which quantify how many individuals move from one compartment to another. In the following, the standard *deterministic* SIR model (Kermack and McKendrick, 1927), which models the average evolution of the epidemic process and does not allow for randomness, is outlined. A comprehensive introduction to the deterministic SIR model and related compartmental models, as well as their stochastic counterparts, is given in the books by Anderson and May (1991), Andersson and Britton (2000) and Keeling and Rohani (2008).

Let $S(t)$, $I(t)$, and $R(t)$ denote the fractions of susceptible, infectious, and recovered individuals, respectively, at time t in a closed population of fixed size. Then

$$S(t) + I(t) + R(t) = 1\,, \quad \text{for all } t.$$

Individuals are first susceptible, they may then get infected and stay infectious for some time before they recover. Initially, there are no recovered individuals and a known fraction of infectious individuals. The standard deterministic SIR model is then defined by the differential equations

$$\begin{aligned} \frac{\mathrm{d}S(t)}{\mathrm{d}t} &= -\beta S(t)I(t) \\ \frac{\mathrm{d}I(t)}{\mathrm{d}t} &= \beta S(t)I(t) - \gamma I(t) \\ \frac{\mathrm{d}R(t)}{\mathrm{d}t} &= \gamma I(t), \end{aligned} \tag{1}$$

where β is the *transmission rate*, and γ is the *recovery rate* with its reciprocal $1/\gamma$ determining the mean infectious period. The crucial term here is $\beta S(t)I(t)$, which states that infections occur at high rate only when there are both many susceptible and many infectious individuals (Andersson and Britton, 2000, Section 1.4).

The potential for an infection to spread within a population is governed by the *basic reproductive ratio* R_0, i.e. the average number of individuals who are directly infected by an initial case in a totally susceptible population (Keeling and Rohani, 2008, p. 20). Rewriting the second equation in (1) as

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = \beta I(t)\left[S(t) - \frac{\gamma}{\beta}\right],$$

it follows that, if the initial proportion of susceptibles $S(0)$ is less than the relative removal rate γ/β, then $\mathrm{d}I(t)/\mathrm{d}t < 0$ and the infection dies out (Kermack and McKendrick, 1927). The inverse of the relative removal rate represents the basic reproductive ratio R_0. Hence, an epidemic can only occur if $R_0 > 1/S(0)$, or if $R_0 > 1$ in an entirely susceptible population with $S(0) = 1$. This threshold property is of paramount importance when investigating the effectiveness and impact of control strategies such as vaccination or school closures.

The above standard SIR model assumes a population of homogeneous individuals who mix, i.e. meet each other, with equal probability. However, this is rarely realistic as there will always be some contacts that are more likely than others. Differences in contacts may be attributed to geographical, behavioral or social factors. Individuals may vary in their susceptibility to infection or in how infectious the individual is when infected or in both. See Becker and Britton (1999) or Grassly and Fraser (2008) for a further discussion of factors influencing disease transmission. In the simplest cases of non-random mixing, the population can be separated into a few homogeneous subpopulations, depending upon certain characteristics that may influence the risk of infection and disease transmission, such as age or gender. In those subpopulations, individuals are similar with respect to both their infectiousness and susceptibility to infection and multitype epidemic models can be used (Britton, 1998). Other extensions of the SIR model which account for further heterogeneity are discussed in Keeling and Rohani (2008).

In practice, infectious disease data are collected at time intervals, rather than continuously. If the infectious period is short compared to the latent period, new infections will occur in

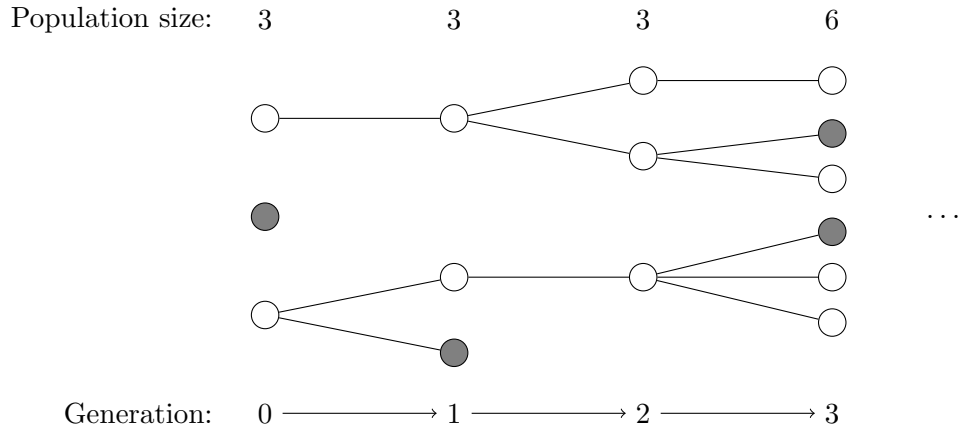


Figure 3.: Schematic illustration of the Galton-Watson branching process. Individuals are represented as bullets, the gray ones produce no offspring.

successive generations and it is natural to consider models in discrete time, where the unit of time corresponds to the *generation time* (Andersson and Britton, 2000, p. 4). See Becker (1989) or Daley and Gani (1999) for a discussion of SIR models in discrete time, so-called chain binomial models.

Branching processes

Information about the number of susceptibles is rarely available, especially for large populations in a surveillance setting. Here, the theory of branching processes proves to be useful for making inference about the reproductive ratio (Becker, 1989, Chapter 8; Andersson and Britton, 2000, Section 3.3). Branching processes are individual-based models for the evolution of populations in terms of generations. They provide an approximation to the epidemic process by assuming an unlimited amount of susceptibles so that the spread of infection depends only on the number of infectious individuals.

Harris (1989) and Haccou, Jagers and Vatutin (2007) provide a comprehensive discussion of general branching processes. In the following, the (Bienaymé-)Galton-Watson process, which represents the oldest and simplest discrete time branching process with non-overlapping generations, is introduced. Figure 3 shows a schematic representation of the process. Let Z_n denote the population size (the number of infectious individuals) in the n-th generation. The process starts with an initial population of Z_0 individuals which produce, independently of each other, a random number of offspring and then die. The total number of offspring determines the population size of the first generation, Z_1. These individuals then produce the second generation, and so on. If the number of offspring of the i-th individual belonging to generation $n-1$ is denoted by X_i, then the population size in generation n is given by (Haccou *et al.*, 2007, p. 14)

$$Z_n = \sum_{i=1}^{Z_{n-1}} X_i,$$

where for each generation the X_i are independently and identically distributed (iid), non-negative and integer-valued random variables. The distribution of X_i is called the *offspring distribution.*

Clearly, once a generation is empty the population dies out. Suppose there is one initial indi-

vidual, $Z_0 = 1$, and the offspring distribution has mean $\lambda = \mathrm{E}(X_i)$. Then the expected size of the n-th generation is given by

$$\mathrm{E}(Z_n) = \mathrm{E}(\mathrm{E}(Z_n|Z_{n-1})) = \lambda\,\mathrm{E}(Z_{n-1}) = \ldots = \lambda^n\,.$$

As $n \to \infty$, the average size of the population goes to zero when $\lambda < 1$, whereas it grows exponentially when $\lambda > 1$ (Haccou *et al.*, 2007, p. 19). In applications to infectious diseases, the offspring mean λ is identified as the reproductive ratio (Becker, 1989, pp. 176f).

The above branching process models the evolution of a closed population. It can be extended to a *branching process with immigration* (Haccou *et al.*, 2007, Section 2.10), where a random number of immigrants, I_n, may arrive during each reproduction period n:

$$Z_n = \sum_{i=1}^{Z_{n-1}} X_i + I_n. \tag{2}$$

Here, I_n and X_i are assumed to be mutually independent and the I_n are iid with finite mean $\nu = \mathrm{E}(I_n)$. Provided that $\lambda < 1$, the process (2) has a stationary distribution with mean

$$\mu = \frac{\nu}{1-\lambda} \tag{3}$$

(Guttorp, 1995, p. 99). Thus, the population cannot ultimately die out due to the immigration.

Statistical inference

As the infection process typically is only partially observed, the detailed data required for estimating the parameters in a mechanistic epidemic model are rarely available. Especially in complex compartmental models allowing for heterogeneity and biological details to achieve greater realism, it is often necessary to assume values for the required parameters and to compare observed data on outbreaks with simulations from the model (Bretó, He, Ionides and King, 2009). The question is how to derive sensible values in the absence of empirical information. Strong modelling assumptions may be necessary and extensive sensitivity analyses are required. The usefulness of such models may be unclear if the model leads to a wide range of predicted effects, and observed data are consistent with many different (and possibly conflicting) interpretations (Cox, Donnelly, Bourne, Gettinby, McInerney, Morrison and Woodroffe, 2005; Cauchemez, Valleron, Boëlle, Flahault and Ferguson, 2008). This emphasizes the need of statistical methods to estimate key parameters of mechanistic models and the associated standard errors based on available data. For a discussion of the statistical challenges inherent to the analysis of infectious diseases, see Becker and Britton (1999) or O'Neill (2002).

Various approaches to make statistical inferences for (complex) mechanistic models based on observed infectious disease time series are proposed in the literature. Much work considers the analysis of measles dynamics, as quite detailed data are available for this disease. For instance, Ellner, Bailey, Bobashev, Gallant, Grenfell and Nychka (1998) suggested a semi-mechanistic SEIR model, where the transmission rate $\beta(t)$ is fitted by a regression model. Finkenstädt and Grenfell (2000) later proposed a discrete-time version of the SIR model that may be fitted to observed time-series of disease incidence in a two-stage approach. First, the unobserved susceptible class is reconstructed using data on births and deaths. Then the transmission equation is fitted using this reconstructed class of susceptibles as a covariate. This model was further developed in Finkenstädt, Bjørnstad and Grenfell (2002) and Xia, Gog and Grenfell (2005). These dis-

crete time models require the epidemic and data collection processes to have similar time scales. Recently, Cauchemez and Ferguson (2008) and Bretó *et al.* (2009) proposed a likelihood-based analysis of continuous time compartmental models. The former obtain maximum likelihood estimates by approximating the SIR epidemic process by a more analytically tractable diffusion process, while the latter employ iterated filtering methods. Further likelihood-based inference approaches for the analysis of infectious diseases other than measles are proposed e.g. by Koelle and Pascual (2004), or Höhle (2009).

Alternatively, Bayesian approaches using MCMC may be used for inference (O'Neill and Roberts, 1999; Morton and Finkenstädt, 2005; Lekone and Finkenstädt, 2006; Forrester, Pettitt and Gibson, 2007; Jewell, Kypraios, Neal and Roberts, 2009). An advantage here is that MCMC methods usually allow the imputation of missing data, such as the unobserved infection times, in a rather straightforward manner. However, analyses may be very time-consuming, especially for large populations and complex models.

An up-to-date overview of existing statistical inference approaches for epidemic models is given by O'Neill (2010). For an account of statistical models based on branching processes see Farrington, Kanaan and Gay (2003) and the references therein.

1.2 Empirical modelling

Empirical models are mainly used for description and prediction of the disease incidence, or in conjunction with monitoring. Typical features of infectious disease data are long-term trends, seasonality and occasional outbreaks (compare Figure 1). First and foremost, a model has to take the epidemic nature, i.e. the inherent dependence of observations, into account. Following Cox (1981), auto-dependencies on the individual level can be represented by either parameter- or observation-driven processes. In parameter-driven models, autocorrelation is introduced through a latent process while in observation-driven models past responses enter directly into the model formulation. Various statistical models are suggested for the analysis of (multivariate) times series of infectious disease counts in the literature. These include, amongst others, generalized linear (mixed) models (GL(M)Ms) and time series approaches.

The applicability of those two model classes for the analysis of surveillance time series data, as presented in Figure 1, is discussed in the following. For this purpose, let y_{it} denote the number of cases observed in unit $i = 1, \ldots, I$ at time point $t = 1, \ldots, T$, where units might represent different age groups or regions. For a specific partition of age, in small areas, or when analyzing rare infectious diseases, the considered time series will contain low numbers of counts. The use of Gaussian models is then not appropriate.

In a regression setting, GL(M)Ms (McCullagh and Nelder, 1989; Fahrmeir and Tutz, 2001) are frequently used to model such non-Gaussian correlated responses. Typically, log-linear Poisson models are applied, where the counts y_{it} are assumed to be Poisson distributed with mean $\mathrm{E}(y_{it})$ which is related to an unknown linear predictor η_{it} via the log link: $\log(\mathrm{E}(y_{it})) = \eta_{it}$. Long-term trends and seasonal variation can easily be incorporated in the linear predictor via linear and cyclic functions of time, for example

$$\eta_{it} = \alpha + \beta t + \sum_{s=1}^{S} (\gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t)), \tag{4}$$

where S is the number of sine and cosine pairs to include and ω_s are Fourier frequencies, e.g. $\omega_s = 2\pi s/52$ for weekly data. This formulation goes back to Serfling (1963) and has since

then been frequently used to model seasonal variation in infectious disease data (e.g. Zeger and Karim, 1991; Le Strat and Carrat, 1999; Held, Höhle and Hofmann, 2005; Nelson and Leroux, 2006). Heterogeneity between units may be addressed by adding suitable covariates, if available, to the right hand side of (4). In the absence of such information, unit-specific random effects may be used (Zeger and Karim, 1991; Kleinman, Lazarus and Platt, 2004).

The above formulation takes regular dependence into account. However, it does not allow for occasional outbreaks. In a parameter-driven formulation, this could be achieved by including time-dependent latent effects that follow an autoregressive structure (Zeger, 1988; Nelson and Leroux, 2006). Alternatively, the number of cases in the past may enter as additional explanatory variable into the model. In the following, such observation-driven formulations are discussed for a univariate time series of cases y_t observed in a specific region i, so the index i is dropped.

Models belonging to the class of autoregressive (integrated) moving average (AR(I)MA) models (Box and Jenkins, 1976; Diggle, 1990) are popular as empirical descriptors for an autocorrelated Gaussian time series. An ARMA model of order (p,q) assumes that the observation at time t is given by a linear combination of p previous observations and q previous random shocks:

$$y_t = \sum_{j=1}^{p} \phi_j y_{t-j} + \epsilon_t + \sum_{k=1}^{q} \theta_k \epsilon_{t-k}\,, \quad \text{with } \epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)\,. \tag{5}$$

Here, $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$ are unknown parameters of the model. Note that, to fit this model, the time series must be stationary. This requires that trends and seasonality are removed from the original data e.g. by differencing, as is done in ARIMA models. Many authors have used this model class to analyze infectious disease surveillance data (e.g Helfenstein, 1986; Watier, Richardson and Hubert, 1991; Allard, 1998; Trottier, Philippe and Roy, 2006). However, ARIMA models are designed for real-valued time series, whereas in a surveillance setting time series of counts are typically available. Especially for rare infectious diseases or diseases in smaller areas, methods for non-negative integer-valued data may be more appropriate.

Various generalizations of AR(MA) models to the non-Gaussian case have been suggested. These may be divided into two classes: a) integer valued AR(MA) models as discrete analogues of the real-valued AR(MA) model (5) and b) GLM-based approaches which combine predictors such as (4) with additional AR(MA) components. See Grunwald, Hyndman, Tedesco and Tweedie (2000), and Kedem and Fokianos (2002, Chapters 4–5) for a comprehensive review. In the following, the main focus is on AR structures. Models of the first class will be called INAR models, whereas models of the second class will be called generalized AR models. Although INAR models have been applied to infectious disease data (Cardinal, Roy and Lambert, 1999), generalized AR models seem to be the more fruitful approach. In INAR models it is not possible to simply remove trends and seasonal variation by transforming the original time series in analogy to Gaussian AR models, because the count data nature needs to be preserved (Enciso-Mora, Neal and Rao, 2009). In contrast, the inclusion of trends and seasonal variation via covariates is natural for generalized AR models.

One can further distinguish generalized AR models depending on whether or not an identity link is used to relate the mean to past responses. A Poisson model with identity link corresponds to an additive model, whereas the use of the log link results in a multiplicative model (see Paper II by Paul, Held and Toschke, 2008). A simple additive AR(1) model is given by

$$y_t | y_{t-1} \sim \text{Po}(\mu_t), \quad \text{with } \mu_t = \lambda y_{t-1} + \nu, \quad \lambda > 0,\, \nu > 0. \tag{6}$$

It corresponds to the autoregressive representation of a branching process with Poisson offspring

and immigration as given in (2) on page 6 (see also Kedem and Fokianos, 2002, Section 5.1.1). Likelihood-based inference for model (6) (with an additional moving average term) is discussed by Fokianos, Rahbek and Tjøstheim (2009). For the analysis of infectious disease surveillance data, Held *et al.* (2005) and Held, Hofmann, Höhle and Schmid (2006) start from model (6) and incorporate features such as seasonality and overdispersion by letting ν vary over time and replacing the Poisson by the negative binomial distribution. Held *et al.* (2005) also discuss an extension to the multivariate case. This multivariate formulation forms the basis of Papers II–IV in this thesis and is further discussed in Section 2.

A simple multiplicative AR(1) model analogous to (6) has mean $\mu_t = \exp(\lambda y_{t-1} + \nu)$. However, this formulation poses some problems as discussed by Zeger and Qaqish (1988). They suggested an alternative model with a modified AR term, which is outlined in Paper II. Further generalizations of this model are proposed by Benjamin, Rigby and Stasinopoulos (2003), who also discuss alternative multiplicative model formulations.

1.3 Surveillance

The systematic ongoing collection and analysis of public health data for the purpose of disease control and prevention is termed surveillance. The scope and values of surveillance are introduced e.g. in M'ikanatha *et al.* (2007). One major goal of surveillance is the timely detection of outbreaks, i.e. aberrations in the space-time pattern of incident cases, as they arise. Over the past decade there has been increasing research interest in developing statistical models for prospective outbreak detection. The critical word here is 'prospective', which poses diverse statistical challenges (Unkel, Farrington, Garthwaite, Robertson and Andrews, 2010). In particular, data from surveillance systems usually suffer from under-reporting and reporting-delays. Thus, detailed mathematical modelling is often not appropriate for public health surveillance and the purpose of early outbreak detection.

In practice, the methods by e.g. Stroup, Williamson and Herndon (1989) and Farrington, Andrews, Beale and Catchpole (1996) or variants thereof are commonly used for routine on-line monitoring. Those methods are based on relatively simple statistical models and the repeated use of confidence intervals to decide on aberrations. Note that the latter implies that only the last time point is used for the decision. In recent years, many surveillance algorithms have been inspired by statistical process control techniques, which accumulate information and account for multiple testing, to detect shifts in mean incidence (Woodall, 2006). For instance, Höhle and Paul (2008) suggest count data regression charts based on the generalized likelihood ratio to monitor changes in the mean incidence of a specific infectious disease. They investigate the early detection of both multiplicative shifts in the mean incidence, as well as the appearance of an additional autoregressive component as in the model by Held *et al.* (2005). Systems that are specifically designed for the surveillance of a single infection may be based on more elaborate models (e.g. Diggle, Rowlingson and Su, 2005; Heisterkamp, Dekkers and Heijne, 2006; Martínez-Beneito, Conesa, López-Quílez and López-Maside, 2008).

For a comprehensive review of statistical methods for detecting aberrations, see Unkel *et al.* (2010), Farrington and Andrews (2004) or Sonesson and Bock (2003). Selected methods in veterinary public health are discussed by Höhle, Paul and Held (2009) in Paper I of this thesis. Many methods consider the univariate case where a single time series is monitored, although surveillance systems usually track more than just one time series. If multiple time series relate to the same underlying process (e.g. one disease stratified by age groups), the consideration of dependencies between time series is likely to be of benefit. The multivariate modelling approach proposed in this thesis is well-suited as basis for multivariate prospective outbreak detection.

For instance, an extension of the univariate regression charts proposed in Höhle and Paul (2008) to such multivariate modelling seems feasible.

2 Model framework

In the following, the basic building block of the multivariate modelling approach, which is used throughout this thesis, is introduced. Further model extensions are summarized in relation to one another, and parameter estimation and model choice are outlined.

2.1 Formulation

Suppose that a specific disease is observed in $i = 1, \ldots, I$ units such as different regions, age groups, or related pathogens. Starting from the univariate branching process formulation in (6), Held *et al.* (2005) suggest the following model for the analysis of multivariate time series of counts: given past data, the counts y_{it} are Poisson distributed with conditional mean

$$\mu_{it} = \lambda y_{i,t-1} + \phi \sum_{j \neq i} y_{j,t-1} + n_{it}\nu_{it}\,, \quad (\lambda, \phi, \nu_{it} > 0) \tag{7}$$

where n_{it} is an offset representing for example the population size of unit i at time t. Here, the first two additive components represent an autoregression on past counts in the same unit i and in other units j. The second component should capture dependency across units as epidemics often spread across regions. The third component is called the 'endemic' component and models regular trend and seasonal patterns in the disease incidence and essentially corresponds to a log-linear regression model as given in (4):

$$\log(\nu_{it}) = \alpha_i + \beta t + \sum_{s=1}^{S}(\gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t)). \tag{8}$$

Note that here, unit-specific intercepts α_i are used to deal with differences in incidence levels. The Poisson distribution for the observed counts may be replaced by a negative binomial distribution to account for overdispersion (Hilbe, 2007, Chapter 6).

Although temporal and spatio-temporal dependence can be accounted for in this model, it may still not be satisfactory in applications. Therefore, further model generalizations may be useful. For instance, the use of $\sum_{j \neq i} y_{j,t-1}$ as explanatory covariate might not be able to adequately capture the spatio-temporal spread of a disease across regions. In Paper II, a weighted sum of past cases is considered, where fixed weights reflect how cases in other regions relate to cases in region i. Such weights may incorporate travel information.

Furthermore, model (7) assumes that disease transmission is equal for all units and also constant over time. However, factors such as age, vaccination status or environmental conditions might influence disease transmission. Changes in vaccination coverage or public health interventions will have an effect on the disease incidence pattern. In Paper II, heterogeneity between units is allowed for by unit-specific parameters, e.g. λ_i or ϕ_i. Such an approach is feasible if the number of units is not too large. In particular, this extension is useful for the joint analysis of several related diseases as different pathogens usually are not equally infectious or might show different seasonality. For instance, there is both clinical and epidemiological evidence that influenza infections predispose meningococcal disease. This is investigated in Paper II based on surveillance

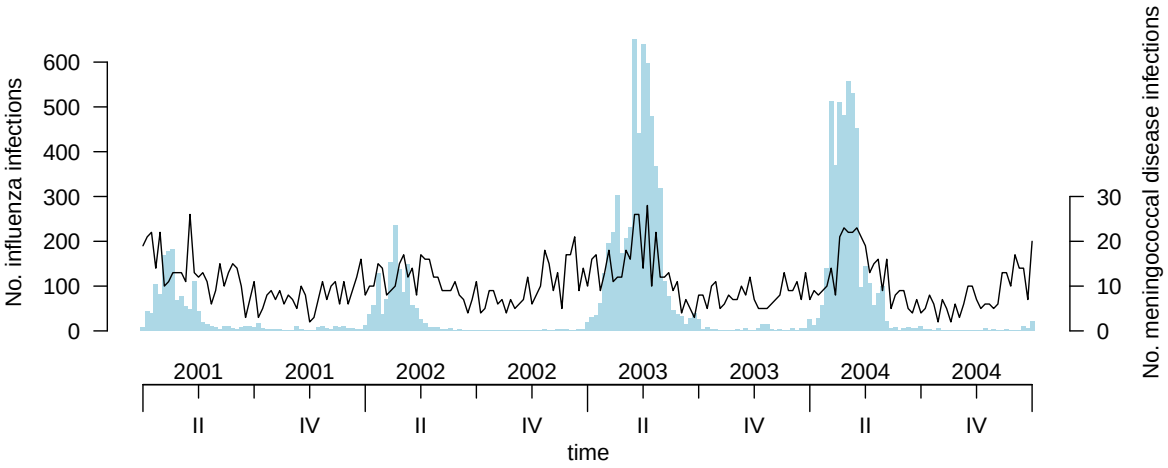


Figure 4.: Number of influenza (blue) and meningococcal disease (black) cases simulated from the second model in Table 2 of Paper II. The autoregressive parameters are specified as $\lambda_{\text{flu}} = 0.74$, and $\lambda_{\text{men}} = 0.10$. The influence of past influenza cases on the meningococcal disease incidence is quantified by $\phi_{\text{flu}\to\text{men}} = 0.005$.

data from Germany. Results of the analysis support this statement although the evidence is not particularly strong. For illustration, Figure 4 displays simulated data from a bivariate model for influenza and meningococcal disease infections, where the number of meningococcal disease cases in week t depends to some extent on the number of influenza cases in the previous week. The realized incidence pattern nicely agrees with the observed incidence as shown in Figure 1 on page 2.

For highly multivariate time series of counts, the use of unit-specific parameters is no longer feasible, as estimation and identifiability problems may occur. Therefore, a random effects formulation is implemented by Paul and Held (2010) in Paper III. Here, random effects may be assumed to be either uncorrelated or spatially correlated. Alternatively, when heterogeneity is due to some known (and observable) factor, this information could be directly related to e.g. the autoregressive parameter via a regression formulation as $\lambda_i = \exp(\boldsymbol{x}^\top \boldsymbol{\beta})$. Herzog, Paul and Held (2010) use region-specific vaccination coverage to explain regional differences in the incidence pattern of measles in Paper IV. With such a regression formulation, it would also be possible to let the autoregressive parameter vary over time.

2.2 Parameter estimation

Statistical inference for model (7) is performed by maximum likelihood (ML) as is done by Held *et al.* (2005). Alternatively, a fully Bayesian approach could have been used for inference, as is suggested in Held *et al.* (2006) for the analysis of univariate time series. However, this requires the use of computer-intensive MCMC methods, whereas results are readily available in a likelihood based approach. The estimation procedure for fitting any model of this thesis was implemented in the `R` package `surveillance`. The procedure is programmed in a general manner and its usage is illustrated in a software tutorial in Appendix III.

The additive decomposition of the mean (7) in combination with (8) implies that the model is non-linear in parameters and does not belong to the class of GLMs. Hence, the Poisson (or negative binomial) log-likelihood $\ell(\boldsymbol{\theta})$ is optimized using suitable optimization routines to obtain ML estimates $\hat{\boldsymbol{\theta}}$. For fixed effects models, ML estimates are obtained by solving the (non-linear)

score equations

$$s(\boldsymbol{\theta}) := \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

e.g. by means of the iterative Newton-Raphson (NR) method which proceeds as follows: Starting with initial estimates $\boldsymbol{\theta}_0$, the next estimates $\boldsymbol{\theta}_t$ are computed as the root of a linear Taylor expansion of the score function around the current estimate $\boldsymbol{\theta}_{t-1}$

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + [F(\boldsymbol{\theta}_{t-1})]^{-1} s(\boldsymbol{\theta}_{t-1}) \,, \quad t = 1, 2, \ldots$$

Here, $F(\boldsymbol{\theta}) := -\partial^2 \ell(\boldsymbol{\theta}) / (\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top)$ is the observed Fisher information matrix. Convergence is reached if the score equation is satisfied, $s(\boldsymbol{\theta}_t) \approx \mathbf{0}$, and differences between successive iterates become small, e.g. $||\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}|| < \epsilon$.

The NR algorithm requires first and second derivatives of the log-likelihood. If analytic derivatives are not available or involve cumbersome computations, numerical approximations may be used. The analytical score function and Fisher information matrix for a general formulation of model (7) are given in Appendix II. Their use considerably improves the time needed to obtain convergence.

The major advantage of the NR method is its rapid convergence. However, good initial values are required even in well-behaved problems, as the algorithm may fail to converge if $\boldsymbol{\theta}_0$ is not sufficiently close to the solution. In particular, the observed Fisher information matrix $F(\boldsymbol{\theta})$ may not be positive definite if $\boldsymbol{\theta}$ is far from $\hat{\boldsymbol{\theta}}$. Dennis and Schnabel (1996, Chapter 6) discuss several modifications of the NR method to obtain a globally convergent method, see also Paper III. For the models considered in this thesis, the optimization routine implemented in the `R` function `nlminb` turned out to be well-suited and more reliable than the optimizers implemented in `optim`.

It is generally advisable to start with simple models that require only few parameters to be estimated. Those estimates can then serve as initial values for parameters in extended models. For instance, Figure 5 displays a contour plot of the estimated log relative likelihood, $\ell(\lambda, \alpha, \hat{\gamma}, \hat{\delta}) - \ell(\hat{\lambda}, \hat{\alpha}, \hat{\gamma}, \hat{\delta})$, for a 4-parameter model with mean

$$\mu_t = \lambda y_{t-1} + \exp\left\{\alpha + \gamma \sin\left(\frac{2\pi}{52} t\right) + \delta \cos\left(\frac{2\pi}{52} t\right)\right\}$$

applied to the meningococcal disease data shown in Figure 1. Note that the autoregressive parameter λ is optimized on log-scale to ensure positive values. Suitable initial values may be obtained by setting the seasonal parameters to zero, $\gamma_0 = \delta_0 = 0$. The plot of the incidence in Figure 1 shows that there is little epidemic behavior and thus a small value for the autoregressive parameter, say $\lambda_0 = 0.1$, is reasonable. Initial values for the intercept α in the endemic component may then be obtained by equating the observed mean $\bar{y}$ to the stationary mean μ in (3) and solving for α.

When random effects are contained in a model, parameter estimation becomes more complicated. The estimation of parameters involves integration of the likelihood with respect to the random effects which cannot be done analytically in non-Gaussian models. Inference in GLMMs is already quite involved, especially if correlated random effects are included. Numerical integration is usually more accurate than approximation methods to the integrand, such as those using first-order Laplace approximations. However, the former is much slower and infeasible if the dimension of the integrals is too large. In Paper III, possible methods for solving the respective integrals are outlined. A penalized likelihood approach is implemented based on methodology

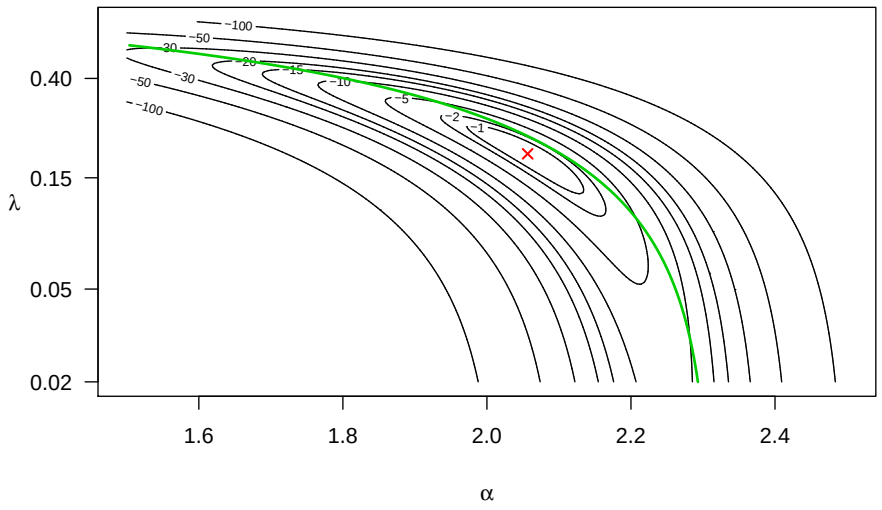


Figure 5.: Contour plot of the estimated log relative likelihood, $\ell(\lambda, \alpha, \hat{\gamma}, \hat{\delta}) - \ell(\hat{\lambda}, \hat{\alpha}, \hat{\gamma}, \hat{\delta})$, for the meningococcal disease data. The red cross shows the ML estimates $(\hat{\alpha}, \hat{\lambda})$, and the green line corresponds to possible initial values obtained by equating $\bar{y} = \exp(\alpha)/(1 - \lambda_0)$ and solving for α given a fixed value λ_0.

for survival data by Kneib and Fahrmeir (2007). Here, estimates of variance components are obtained by optimizing an approximated marginal likelihood which is obtained via a first-order Laplace approximation. The estimation procedure is adapted to handle correlated random effects. See Appendix II for further details.

2.3 Model choice

Model choice is a fundamental part of data analysis. It includes the selection of relevant covariates, random effects, or the distribution of the response. Classical model choice criteria such as Akaike's information criterion (Akaike, 1974, AIC) or the Bayesian information criterion (Schwarz, 1978, BIC) are widely used. Both criteria combine a measure for model fit with a penalty for model complexity. The AIC is defined by

$$\text{AIC} = -2\log(L) + 2p\,,$$

where $\log(L)$ is the maximized log-likelihood and p denotes the number of estimable parameters in the considered model. The BIC replaces the penalty term by $p\log(n)$, where n denotes the sample size. The model with the smallest AIC (BIC) is considered the best.

While model selection based on such information criteria is well explored and understood for models that correspond to fixed-effects likelihoods, their use turns out to be problematic once models include random effects (Burnham and Anderson, 2002, p. 316). The arising difficulties are outlined in Paper III.

For model selection in time series models, the comparison of successive one-step-ahead forecasts with the actually observed data is suggested (Dawid, 1984). In this context, the use of strictly proper scoring rules, such as the logarithmic score or the ranked probability score, is recommended (Gneiting and Raftery, 2007; Czado, Gneiting and Held, 2009). These criteria are described and used for model selection for the random effects models in Paper III.

Thesis Summary

This thesis consists of four papers, presented in chronological order. Their content is briefly summarized in the following:

Paper I

Paper I, **Statistical approaches to the monitoring and surveillance of infectious diseases for veterinary public health** by Michael Höhle, Michaela Paul and Leonhard Held, provides an overview of methodology for the prospective and retrospective surveillance of routinely collected count data in veterinary public health. Particular emphasis is placed on the time series nature of the data. Various methods for the prospective detection of aberrations are discussed, most of which use statistical process control techniques. Statistical modelling approaches such as the model by Held *et al.* (2005) allow a retrospective investigation of temporal and spatio-temporal patterns in the data. The applicability of the presented methods is illustrated with data from the monitoring of rabies among fox in the federal state of Hesse, Germany.

This work stems from a joint talk by L. Held and M. Höhle at the GisVet 2007 conference in Copenhagen. M. Höhle was the main investigator and drafted the manuscript. I performed the computations for the retrospective analysis of the rabies data using the fitting procedure implemented by me in preparation for Paper II. Together, L. Held and I wrote the section about this retrospective analysis and commented on the other parts of the manuscript. The paper was finalized by M. Höhle.

The main contribution of this paper is the description of powerful statistical methods for univariate and hierarchical surveillance using the example of rabies monitoring. The existence of a readily available implementation of these methods in the `R` package `surveillance` (Höhle, 2007) facilitates their use in practice.

Paper II

In Paper II, **Multivariate modelling of infectious disease surveillance counts** by Michaela Paul, Leonhard Held and André M. Toschke, the model proposed in Held *et al.* (2005) is extended to analyze (directed or directionless) dependencies between related diseases observed in one area, and dependencies between a specific disease observed in several areas. The method is illustrated through a joint analysis of weekly influenza and meningococcal disease counts in Germany. In a second application, air travel information is included in the model to analyze the spatio-temporal spread of influenza in the USA. Model choice is based on AIC.

This work is based on initial work in the Master's thesis of Toschke (2005), supervised by L. Held. The model formulation and estimation approach were refined by L. Held and me. I implemented and integrated the extended fitting procedure in the `R` package `surveillance`. In consultation with L. Held, I conducted all analyses and did most of the writing. A. M. Toschke commented on the manuscript which was finalized by L. Held and me.

The main contribution of this work is to provide a flexible model-based approach for the analysis of multivariate time series data on counts of infectious diseases. Different types of variation and correlation can be incorporated within a single model. For instance, it is possible to explore potentiating effects of one infection on another, or directionless associations between infections transmitted via the same route which might be induced by varying contact rates. A comparison

with the multiplicative Markov regression model by Zeger and Qaqish (1988) shows empirical evidence of a better fit of the proposed additive model formulation.

Paper III

In Paper III, **Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts** by Michaela Paul and Leonhard Held, region-specific random effects are introduced in the model by Paul *et al.* (2008) to account for heterogeneous incidence levels or varying disease transmission. The random effects are assumed to be uncorrelated or spatially correlated. Inference is based on penalized likelihood methodology for mixed models (Breslow and Clayton, 1993; Kneib and Fahrmeir, 2007). In a case study, the model is applied to monthly counts of meningococcal disease in 94 departments of France and weekly counts of influenza cases in 140 districts of Southern Germany.

This work extends Paper II to analyze surveillance data reported in a large number of regions. After discussion of the inference approach with L. Held, I worked out the methodological details, implemented the fitting procedure, conducted all analyses and wrote a draft of the manuscript. L. Held commented on the draft, which I finalized.

The main contribution of this paper is the formulation via (possibly spatially correlated) random effects. This enables a realistic analysis of a large number of parallel time series, where heterogeneity and correlations across regions are very likely to exist. Random effects may enter into each of the three model components, and correlations are allowed both across regions as well as across components. Classical model choice criteria such as AIC or BIC can be problematic in the presence of random effects. Thus models are compared by means of one-step-ahead predictions and proper scoring rules. As exemplified by the applications, the predictive performance improves, if existing heterogeneity is accounted for using random effects.

For further information, Appendix I presents an early version of Paper III which is published in the proceedings of the 16^{th} European Young Statisticians meeting in Bucharest, 24–28th August 2009. In this version, the special case of uncorrelated random effects in a model formulation without the spatio-temporal component is discussed. The diagonal structure of the random effects covariance matrix simplifies matters compared to the more general formulation in Paper III. The paper also contains preliminary results for the French meningococcal disease data.

Paper IV

In Paper IV, **Heterogeneity in vaccination coverage explains the size and occurrence of measles epidemics in German surveillance data** by Sereina A. Herzog, Michaela Paul and Leonhard Held, the link between local measles outbreaks in Germany and vaccination coverage levels is characterized, given that most cases during such outbreaks occurred in unvaccinated people. This is investigated by introducing region-specific vaccination coverage levels as explanatory variable in the multivariate time-series model proposed by Held *et al.* (2005). Model choice is based on AIC, the behavior of which is investigated in a simulation study.

This work originated from an idea of L. Held which was initially explored in the Master's thesis of Herzog (2008), supervised by L. Held and in part by me. S. Herzog wrote a first draft based on her thesis. After discussion with L. Held and S. Herzog, I revised the manuscript in a second draft. S. Herzog carried out the simulation study, I conducted the analyses of the measles data using `R` functions from S. Herzog's Master's thesis. Together, we finalized the paper, with L. Held giving comments.

Heterogeneity in diseases transmitted by person-to-person contact can be caused by various factors such as vaccination status, age, genetic variation in individuals or environmental conditions. Compared to Papers II and III the main contribution of Paper IV is to present an alternative approach to address heterogeneity in disease incidence and transmission when suitable covariate information about such factors is available. Results of the analysis of German measles surveillance data show a clear association between vaccination coverage and the overall incidence of measles in the federal states of Germany.

Appendix II presents details about the estimation procedure for the most general formulation of the model including correlated random effects as presented in Paper III. Additional covariates as discussed in Paper IV may also be included in the model formulation.

The estimation procedure was implemented by me in a general manner, and is integrated in the `R` package `surveillance`. Appendix III presents a software tutorial to illustrate its usage.

References

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **19**(6): 716–723.

Allard, R. (1998). Use of time-series analysis in infectious disease surveillance, *Bulletin of the World Health Organization* **76**(4): 327–333.

Anderson, R. M. and May, R. M. (1991). *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press, Oxford.

Andersson, H. and Britton, T. (2000). *Stochastic Epidemic Models and Their Statistical Analysis*, Lecture Notes in Statistics 151, Springer, New York.

Becker, N. G. (1989). *Analysis of Infectious Disease Data*, Monographs on Statistics and Applied Probability, Chapman & Hall, London.

Becker, N. G. and Britton, T. (1999). Statistical studies of infectious disease incidence, *Journal of the Royal Statistical Society. Series B* **61**(2): 287–307.

Benjamin, M. A., Rigby, R. A. and Stasinopoulos, D. M. (2003). Generalized autoregressive moving average models, *Journal of the American Statistical Association* **98**(461): 214–223.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*, revised edn, Holden-Day, San Francisco, Calif.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**(421): 9–25.

Bretó, C., He, D. H., Ionides, E. L. and King, A. A. (2009). Time series analysis via mechanistic models, *Annals of Applied Statistics* **3**(1): 319–348.

Britton, T. (1998). Estimation in multitype epidemics, *Journal of the Royal Statistical Society. Series B* **60**(4): 663–679.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach*, 2nd edn, Springer, New York.

Cardinal, M., Roy, R. and Lambert, J. (1999). On the application of integer-valued time series models for the analysis of disease incidence, *Statistics in Medicine* **18**(15): 2025–2039.

Cauchemez, S. and Ferguson, N. M. (2008). Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London, *Journal of the Royal Society Interface* **5**(25): 885–897.

Cauchemez, S., Valleron, A. J., Boëlle, P. Y., Flahault, A. and Ferguson, N. M. (2008). Estimating the impact of school closure on influenza transmission from sentinel data, *Nature* **452**(7188): 750–755.

Cox, D. (1981). Statistical analysis of time series. Some recent developments, *Scandinavian Journal of Statistics* **8**(2): 93–115.

Cox, D. R., Donnelly, C. A., Bourne, F. J., Gettinby, G., McInerney, J. P., Morrison, W. I. and Woodroffe, R. (2005). Simple model for tuberculosis in cattle and badgers, *Proceedings of the National Academy of Sciences of the United States of America* **102**(49): 17588–17593.

Czado, C., Gneiting, T. and Held, L. (2009). Predictive model assessment for count data, *Biometrics* **65**(4): 1254–1261.

Daley, D. J. and Gani, J. (1999). *Epidemic Modelling: An Introduction*, Cambridge University Press, Cambridge.

Dawid, A. P. (1984). Statistical theory. The prequential approach, *Journal of the Royal Statistical Society. Series A* **147**(2): 278–292.

Dennis, J. E. and Schnabel, R. B. (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Society for Industrial and Applied Mathematics, Philadelphia.

Diggle, P. J. (1990). *Time Series. A Biostatistical Introduction*, Oxford University Press, Oxford.

Diggle, P. J., Rowlingson, B. and Su, T.-L. (2005). Point process methodology for on-line spatio-temporal disease surveillance, *Environmetrics* **16**(5): 423–434.

Ellner, S. P., Bailey, B. A., Bobashev, G. V., Gallant, A. R., Grenfell, B. T. and Nychka, D. W. (1998). Noise and nonlinearity in measles epidemics: combining mechanistic and statistical approaches to population modeling, *American Naturalist* **151**(5): 425–440.

Enciso-Mora, V., Neal, P. and Rao, T. S. (2009). Integer valued AR processes with explanatory variables, *Sankhyā: The Indian Journal of Statistics* **71-B**(2): 248–263.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd edn, Springer, Berlin.

Farrington, C. P. and Andrews, N. (2004). Outbreak detection: application to infectious disease surveillance, *in* R. Brookmeyer and D. F. Stroup (eds), *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance*, Oxford University Press, New York, pp. 233–266.

Farrington, C. P., Andrews, N. J., Beale, A. D. and Catchpole, M. A. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease, *Journal of the Royal Statistical Society. Series A* **159**(3): 547–563.

Farrington, C. P., Kanaan, M. N. and Gay, N. J. (2003). Branching process models for surveillance of infectious diseases controlled by mass vaccination, *Biostatistics* **4**(2): 279–295.

Finkenstädt, B. F. and Grenfell, B. T. (2000). Time series modelling of childhood diseases: a dynamical systems approach, *Journal of the Royal Statistical Society. Series C* **49**(2): 187–205.

Finkenstädt, B. F., Bjørnstad, O. N. and Grenfell, B. T. (2002). A stochastic model for extinction and recurrence of epidemics: estimation and inference for measles outbreaks, *Biostatistics* **3**(4): 493–510.

Fokianos, K., Rahbek, A. and Tjøstheim, D. (2009). Poisson autoregression, *Journal of the American Statistical Association* **104**(488): 1430–1439.

Forrester, M. L., Pettitt, A. N. and Gibson, G. J. (2007). Bayesian inference of hospital-acquired infectious diseases and control measures given imperfect surveillance data, *Biostatistics* **8**(2): 383–401.

Giesecke, J. (2002). *Modern Infectious Disease Epidemiology*, 2 edn, Hodder Arnold, London.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association* **102**(477): 359–378.

Grassly, N. C. and Fraser, C. (2008). Mathematical models of infectious disease transmission, *Nature Reviews Microbiology* **6**(6): 477–487.

Grunwald, G. K., Hyndman, R. J., Tedesco, L. and Tweedie, R. L. (2000). Non-Gaussian conditional linear AR(1) models, *Australian & New Zealand Journal of Statistics* **42**(4): 479–495.

Guttorp, P. (1995). *Stochastic Modelling of Scientific Data*, Chapman and Hall, London.

Haccou, P., Jagers, P. and Vatutin, V. A. (2007). *Branching Processes: Variation, Growth, and Extinction of Populations*, Cambridge University Press, Cambridge.

Harris, T. E. (1989). *The Theory of Branching Processes*, Dover Publications, New York. Corrected reprint of the 1963 original (Springer, Berlin).

Heisterkamp, S. H., Dekkers, A. L. M. and Heijne, J. C. M. (2006). Automated detection of infectious disease outbreaks: hierarchical time series models, *Statistics in Medicine* **25**(24): 4179–4196.

Held, L., Hofmann, M., Höhle, M. and Schmid, V. (2006). A two-component model for counts of infectious diseases, *Biostatistics* **7**(3): 422–437.

Held, L., Höhle, M. and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts, *Statistical Modelling* **5**(3): 187–199.

Helfenstein, U. (1986). Box-Jenkins modeling of some viral infectious diseases, *Statistics in Medicine* **5**(1): 37–47.

Herzog, S. A. (2008). *Regression modelling of counts of communicable diseases with time-varying infectiousness*, Master's thesis, University of Zurich.

Herzog, S. A., Paul, M. and Held, L. (2010). Heterogeneity in vaccination coverage explains the size and occurrence of measles epidemics in German surveillance data, *Epidemiology and Infection*. doi:10.1017/S0950268810001664.

Hilbe, J. M. (2007). *Negative Binomial Regression*, Cambridge University Press, Cambridge.

Höhle, M. (2007). Surveillance: an R package for the monitoring of infectious diseases, *Computational Statistics* **22**(4): 571–582.

Höhle, M. (2009). Additive-multiplicative regression models for spatio-temporal epidemics, *Biometrical Journal* **51**(6): 961–978.

Höhle, M. and Paul, M. (2008). Count data regression charts for the monitoring of surveillance time series, *Computational Statistics and Data Analysis* **52**(9): 4357–4368.

Höhle, M., Paul, M. and Held, L. (2009). Statistical approaches to the monitoring and surveillance of infectious diseases for veterinary public health, *Preventive Veterinary Medicine* **91**(1): 2–10.

Jewell, C. P., Kypraios, T., Neal, P. and Roberts, G. O. (2009). Bayesian analysis for emerging infectious diseases, *Bayesian Analysis* **4**(3): 465–469.

Kedem, B. and Fokianos, K. (2002). *Regression Models for Time Series Analysis*, Wiley-Interscience, Hoboken, NJ.

Keeling, M. J. and Rohani, P. (2008). *Modeling Infectious Diseases in Humans and Animals*, Princeton University Press, Princeton, NJ.

Kermack, W. O. and McKendrick, A. G. (1927). Contributions to the mathematical theory of epidemics, *Proceedings of the Royal Society of London Series A* **115**(772): 700–721.

Kleinman, K., Lazarus, R. and Platt, R. (2004). A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism, *American Journal of Epidemiology* **159**(3): 217–224.

Kneib, T. and Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression, *Scandinavian Journal of Statistics* **34**(1): 207–228.

Koelle, K. and Pascual, M. (2004). Disentangling extrinsic from intrinsic factors in disease dynamics: a nonlinear time series approach with an application to cholera, *American Naturalist* **163**(6): 901–913.

Le Strat, Y. and Carrat, F. (1999). Monitoring epidemiologic surveillance data using hidden Markov models, *Statistics in Medicine* **18**(24): 3463–3478.

Lekone, P. E. and Finkenstädt, B. F. (2006). Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study, *Biometrics* **62**(4): 1170–1177.

Martínez-Beneito, M. A., Conesa, D., López-Quílez, A. and López-Maside, A. (2008). Bayesian Markov switching models for the early detection of influenza epidemics, *Statistics in Medicine* **27**(22): 4455–4468.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, 2nd edn, Chapman & Hall, London.

M'ikanatha, N. M., Lynfield, R., Julian, K. G., Van Beneden, C. A. and de Valk, H. (2007). Infectious disease surveillance: a cornerstone for prevention and control, *in* N. M. M'ikanatha, R. Lynfield, H. de Valk and C. A. Van Beneden (eds), *Infectious Disease Surveillance*, Blackwell, Malden, pp. 3–17.

Morton, A. and Finkenstädt, B. F. (2005). Discrete time modelling of disease incidence time series by using Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society. Series C* **54**(3): 575–594.

Nelson, K. P. and Leroux, B. G. (2006). Statistical models for autocorrelated count data, *Statistics in Medicine* **25**(8): 1413–1430.

O'Neill, P. D. (2002). A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods, *Mathematical Biosciences* **180**(1-2): 103–114.

O'Neill, P. D. (2010). Introduction and snapshot review: Relating infectious disease transmission models to data, *Statistics in Medicine* **29**(20): 2069–2077.

O'Neill, P. D. and Roberts, G. O. (1999). Bayesian inference for partially observed stochastic epidemics, *Journal of the Royal Statistical Society. Series A* **162**(1): 121–129.

Paul, M. and Held, L. (2009). Incorporating random effects in a likelihood-based model for multivariate infectious disease counts, *Proceedings of the* 16^{th} *European Young Statisticians meeting*, Bucharest, Romania.

Paul, M. and Held, L. (2010). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Statistics in Medicine*. Accepted.

Paul, M., Held, L. and Toschke, A. M. (2008). Multivariate modelling of infectious disease surveillance data, *Statistics in Medicine* **27**(29): 6250–6267.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford University Press, Oxford.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. `http://www.R-project.org`.

Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics* **6**(2): 461–464.

Serfling, R. (1963). Methods for current statistical analysis of excess pneumonia-influenza deaths, *Public Health Reports* **78**(6): 494–506.

Sonesson, C. and Bock, D. (2003). A review and discussion of prospective statistical surveillance in public health, *Journal of the Royal Statistical Society. Series A* **166**(1): 5–21.

Stroup, D. F., Williamson, G. D. and Herndon, J. L. (1989). Detection of aberrations in the occurrence of notifiable diseases surveillance data, *Statistics in Medicine* **8**(3): 323–329.

Toschke, A. M. (2005). *Multivariate modelling of infectious disease surveillance data*, Master's thesis, Ludwig-Maximilians University, Munich.

Trottier, H., Philippe, P. and Roy, R. (2006). Stochastic modeling of empirical time series of childhood infectious diseases data before and after mass vaccination, *Emerging Themes in Epidemiology* **3**: 9.

Unkel, S., Farrington, C. P., Garthwaite, P. H., Robertson, C. and Andrews, N. (2010). A review of statistical methods for the detection of infectious disease outbreaks, *Technical report 10/08*, The Open University Statistics Group. Available from: `http://stats-www.open.ac.uk/TechnicalReports/OutbreakReviewPaper.pdf`.

Watier, L., Richardson, S. and Hubert, B. (1991). A time-series construction of an alert threshold with application to S. bovismorbificans in France, *Statistics in Medicine* **10**(10): 1493–1509.

Woodall, W. H. (2006). The use of control charts in health-care and public-health surveillance, *Journal of Quality Technology* **38**(2): 89–104.

Xia, Y. C., Gog, J. R. and Grenfell, B. T. (2005). Semiparametric estimation of the duration of immunity from infectious disease time series: influenza as a case-study, *Journal of the Royal Statistical Society. Series C* **54**(3): 659–672.

Zeger, S. L. (1988). A regression model for time-series of counts, *Biometrika* **75**(4): 621–629.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects: a Gibbs sampling approach, *Journal of the American Statistical Association* **86**(413): 79–86.

Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach, *Biometrics* **44**(4): 1019–1031.

PAPER I

Statistical approaches to the monitoring and surveillance of infectious diseases for veterinary public health

Michael Höhle, Michaela Paul & Leonhard Held

Paper published in *Preventive Veterinary Medicine* 2009, **91**, 2–10.

Preventive Veterinary Medicine 91 (2009) 2–10

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed

Statistical approaches to the monitoring and surveillance of infectious diseases for veterinary public health☆

Michael Höhle [a,*], Michaela Paul [b], Leonhard Held [b]

[a] Department of Statistics, University of Munich, Ludwigstr. 33, 80539 Munich, Germany
[b] Biostatistics Unit, Institute of Social and Preventive Medicine, University of Zurich, Switzerland

ARTICLE INFO

Keywords:
Veterinary public health
Outbreak detection
Rabies
GIS

ABSTRACT

This paper covers the aspect of using statistical methodology for the monitoring and surveillance of routinely collected data in veterinary public health. An account of the Farrington algorithm and Poisson cumulative sum schemes for the prospective detection of aberrations is given with special attention devoted to the occurrence of seasonality and spatial aggregation of the time series. Modelling approaches for retrospective analysis of surveillance counts are also described. To illustrate the applicability of the methodology in veterinary public health, data from the monitoring of rabies among fox in Hesse, Germany, are analysed.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

The specific aim of disease monitoring and surveillance, which has a long history in veterinary sciences as described in e.g. Wilkinson (1992), is the early detection of emerging and re-emerging diseases in order to prepare contingency plans to contain those diseases. Following Christensen (2001) we shall differentiate between *animal disease monitoring* as the ongoing efforts directed at assessing the health and disease status of a given population and *disease surveillance*, which describes a more active system and implies that some form of directed action will be taken if the data indicate aberrations in disease level. An example of monitoring could be the process of keeping track of a known disease, e.g. foot an mouth disease, classical swine fever or rabies. This is in contrast to the detection of non-specific signals as part of early detection for specific infections not present in the population. Such surveillance could, e.g. be the monitoring of mortality in poultry triggering sampling and testing for highly pathogenic avian influenza in the affected poultry flocks. However, as described in Doherr and Audigé (2001) it is common to use the term disease monitoring and surveillance system (MOSS) as umbrella term for the two activities and we shall adopt this terminology in what follows.

The focus of this paper is on statistical methodology for performing prospective outbreak detection and retrospective modelling in time series of disease counts resulting from continuous data collection within a disease MOSS. This task has become increasingly important as the amount of data gathered through automatic data collection increases. For a general overview on statistical challenges for survey design and diagnostic testing in veterinary MOS systems, see Salman (2003).

Examples from human epidemiology include the monitoring of notifiable diseases, congenital malformations, surgical outcomes and bioterrorism syndromes (Widdowson et al., 2003; Chen, 1978; Steiner et al., 2000; Bravata et al., 2004). Examples in veterinary epidemiology are the monitoring of salmonella in livestock reports (Kosmider et al., 2006) or abortions in dairy cattle (Carpenter et al., 2007). One issue in adapting statistical outbreak detection methods from humans to animals is the fundamental differences in terms: Animal MOS systems have to deal with diverse species and completely different

☆ This paper is part of a special issue entitled "GisVet 2007", Guest Edited by Annette Kjær Ersbøll.

* Corresponding author.
E-mail address: hoehle@stat.uni-muenchen.de (M. Höhle).

0167-5877/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.prevetmed.2009.05.017

living conditions (e.g. production, wild, or companion animal) resulting in different entities of interest (e.g. individual or herd) and different professionals interacting with the population. As a consequence the possibility and cost of investigation and control strategy depends heavily on character of living and the mobility of the animal population. However, as much as the human and veterinary MOS systems differ, zoonoses like salmonellosis, rabies or emerging zoonoses (e.g. avian flu) underline the need for a comparative and co-operative approach in monitoring and surveillance.

Data quality is a major practical concern in the analysis of MOSS data, which complicates the statistical analysis. Examples are the lack of a clear case definition, imperfect diagnostic tests, under-reporting, reporting delays and reporting of only test positives with no information on the total number of tests conducted (lack of denominator data). In this paper the focus is however on the statistical challenges of analysing the resulting univariate and multivariate time series containing daily, weekly or monthly counts. This task shall – following Lawson and Kleinman (2005) – be denoted as statistical surveillance of count data time series.

The first part of the article deals with prospective statistical surveillance. Several known methods from the literature are treated, but also new methodological developments such as weighting timeliness of data is presented. One important question in the MOSS is deciding on the best level of aggregation. To this end a new scheme for hierarchical disease detection is introduced. Our presentation thus emphasizes the time series nature of the data as an alternative to spatial and spatio-temporal cluster detection methods, e.g. scan statistics (Kulldorff, 2001; Rogerson, 2001) or point-process oriented approaches (Diggle et al., 2005). The second part of the text describes a stochastic model for the analysis of multivariate surveillance data. This model can be used to detect temporal and spatio-temporal dependencies in multivariate time series of MOSS counts. Applicability of the presented methods is illustrated throughout the text with data from the monitoring database of the WHO rabies surveillance program (WHO Collaboration Centre for Rabies Surveillance and Research, 2007).

2. Prospective surveillance

In this section, statistical methods for univariate and hierarchical disease surveillance are discussed with a focus on outbreak detection for count data with seasonality. Broader surveys of outbreak detection methods can be found in Farrington and Andrews (2003); Sonesson and Bock (2003); Lawson and Kleinman (2005); Buckeridge et al. (2005).

2.1. Univariate surveillance

For many surveillance problems univariate time series of counts are readily available. If not additional preprocessing is performed, e.g. by aggregating geo-referenced outbreak data to an appropriate level or by time-wise aggregation of event time data. We denote the resulting univariate time series by $\{y_t; t = 1, 2, \ldots\}$. Prospective outbreak detection can be seen as a classification task: based on the observed values $y_1, \ldots, y_n$ it is to be decided if there is an aberration at time n or not. In what follows two classes of methods that address the problem are described.

2.1.1. Farrington method

The core of the method by Farrington et al. (1996) is to predict the observed value y_n using a set of reference values taken from the observed values $y_1, \ldots, y_{n-1}$. To handle long-term trends and seasonality, only values from a window of size $2w + 1$ around time n upto b years back in time are taken. Thus, the set of reference values consists of recent values with similar conditions as at time n and can formally be defined as

$$R(w, b) = \left(\bigcup_{i=1}^{b} \bigcup_{j=-w}^{w} y_{n-i\cdot r+j} \right),$$

where r is the period of the observations, e.g. for monthly data r is 12. Thus no observations from the current year are used. Poisson regression with overdispersion is then used to model the $(2w + 1)b$ reference values, i.e. for $y_t \in R(w, b)$

$$\mathrm{E}(y_t) = \mu_t, \quad \text{with} \quad \log \mu_t = \alpha + \beta t \quad \text{and} \quad \mathrm{Var}(y_t) = \phi \mu_t.$$

Based on the estimated model a one-sided $(1 - \kappa) \times 100\%$ prediction interval for y_n can be formed. The classical way to compute such a prediction interval is based on the normal distribution, however, as the skewness of the Poisson distribution with mean μ is $1/\sqrt{\mu}$, for low valued μ this is a bad approximation. Therefore a (2/3)-power transformation is applied to normalize the distribution before computing the interval. The resulting back-transformed upper limit of the prediction interval for y_n is then

$$U_n = \hat{\mu}_n \left\{ 1 + \frac{2}{3} z_{1-\kappa} \cdot \sqrt{\frac{\hat{\phi}\hat{\mu}_n + \mathrm{Var}(\hat{\mu}_n)}{\hat{\mu}_n^2}} \right\}^{3/2},$$

where $\hat{\mu}_n = \exp(\hat{\alpha} + n\hat{\beta})$ and $z_{1-\kappa}$ is the $100(1 - \kappa)\%$ quantile of the standard normal distribution. Subsequently, if $y_n > U_n$ an alarm is sounded. To ease exposition some details of the algorithm have been left out in the above description; e.g. the linear trend is only included if it is significant at the 5% level and a second round of estimation is performed with observations weighted by their inverse residuals. The latter corrects for possible past outbreaks in the reference values. Furthermore, protection against preposterous alarms is made by post-processing alarms and only reporting those where enough cases have been seen.

One virtue of the Farrington method is its simple yet flexible modelling depending on only one user specified parameter κ. Hence, virtually no time series specific tuning is required, which becomes advantageous when applying the method to multiple surveillance time series. One shortcoming is that only a moving window of historical values is taken for estimation with no values taken from the current year. A simple extension would be to include

seasonal terms into the linear predictor of the Poisson regression

$$\log \mu_t = \alpha + \beta t + \sum_{s=1}^{S}\left[\gamma_s \sin\left(\frac{2\pi}{r} s \cdot t\right) + \delta_s \cos\left(\frac{2\pi}{r} s \cdot t\right)\right], \quad (1)$$

where r is the known period. Reference values could then consist of all historical values or all values within a moving window of b years. However, sequential estimation becomes more complicated and careful selection of the parameter S is required. Such modelling is thus only of interest for time series where performance is vital or where additional information is immediately available in order to extend (1) with further covariates.

Another shortcoming is the recomputing of prediction-intervals at every time instance. As a consequence, sustained shifts become hard to detect as deviations are not accumulated. Performance on such shifts can be improved by exploiting methods from statistical process control (SPC), which accumulate information. The cumulative sum (CUSUM) scheme described in the following section is such a method.

2.1.2. CUSUM likelihood ratio detectors

Statistical process control has its root in the quality control of manufactured goods, but has since found numerous application in health care (Woodall, 2006) by providing a more formal setting for surveillance methods. In our treatment we shall put special emphasis on the count data character of the time series – a feature which is less typical for SPC methods.

Central in SPC is the prospective detection of change-points. With n observations and known change-point τ one assumes the following model:

$$y_t|z_t, \tau \sim \begin{cases} f_{\theta_0}(\cdot|z_t) & \text{for } t = 1, \ldots, \tau - 1 \text{(in-control)} \\ f_{\theta_1}(\cdot|z_t) & \text{for } t = \tau, \tau + 1, \ldots \text{(out-of-control)} \end{cases}$$

where z_t denotes known covariates at time t and f_θ is, e.g. the Poisson probability function with its mean μ being a function of θ and z_t. Objective of prospective change-point detection is to use the observations $y_1, \ldots, y_n$ to decide at time n whether a change-point has occurred during $1, \ldots, n$. One approach to do this is to use the so called cumulative sum likelihood ratio detector (Lai, 1995; Frisén, 2003)

$$N = \min\left\{n \geq 1 : \max_{1 \leq \tau \leq n}\left[\sum_{t=\tau}^{n} \log\left\{\frac{f_{\theta_1}(y_t|z_t)}{f_{\theta_0}(y_t|z_t)}\right\}\right] \geq c\right\}. \quad (2)$$

For a specific n the above detector computes the log likelihood ratio (LR) statistic for testing the hypothesis that all observations originate from the in-control distribution against the alternative that from change-point τ on they stem from the out-of-control distribution. Maximizing the LR statistic for each possible change-point $1 \leq \tau \leq n$ means finding the maximum likelihood estimator, $\hat{\tau}$, for the most likely location of the change-point. If $\mathrm{LR}(\hat{\tau})$ is above a pre-specified threshold c, then there is enough information at time n to say that a change-point happened at $\hat{\tau}$. Otherwise, no decision is made and the monitoring continues at time $n + 1$.

Without covariates and with pre-specified θ_0 and θ_1 the above detector can be written in the well known CUSUM recursive form

$$l_0 = 0, \quad l_n = \max\left(0, l_{n-1} + \log\left\{\frac{f_{\theta_1}(y_n)}{f_{\theta_0}(y_n)}\right\}\right), \quad n \geq 1 \quad (3)$$

where the first alarm is given at time $N = \min\{n : l_n \geq c\}$. This detector can be shown to be optimal (in some technical sense) for the detection of a shift from θ_0 to θ_1. Evaluating the performance of the proposed schemes is a question about which performance criterion to consider. Typical choices are location parameters of the run length distribution, e.g. the average run lengths (ARLs) $\mathrm{ARL}_0 = \mathrm{E}(N|\tau = \infty)$, i.e. expected waiting time until the first false alarm, or $\mathrm{ARL}_1 = \mathrm{E}(N|\tau = 0)$, i.e. expected time until detection of the change, when this change occurs immediately. Alternatives include the conditional expected delay $E(N - \tau|\tau, N \geq \tau)$ or the probability of false alarm within the first m time points $P(N \leq m|\tau = \infty)$. Frisén (2003) gives a thorough treatment of the various criteria and available optimality results. One can also use the classification framework to discuss performance using sensitivities, specificities and ROC curves (Kleinmann and Abrams, 2006).

Lucas (1985) covers the CUSUM likelihood ratio method for the Poisson distribution with constant parameters $\mu_0 = \theta_0$ and $\mu_1 = \theta_1$. In a surveillance context main interest is in detecting upward changes, thus typically $\theta_1 > \theta_0$. Dividing by an appropriate constant in (3) one obtains for this upward detection the equivalent form

$$S_n = \max(0, S_{n-1} + (y_n - k)),$$

where $k = (\mu_1 - \mu_0)/(\log(\mu_1) - \log(\mu_0))$ and $S_0 = 0$. A common way to select μ_0 is to use a period known to be in-control to estimate μ_0. In addition, μ_1 is the change, which is to be detected quickly. Note that if the in-control assumption in the μ_0 estimation is violated, e.g. if the training data contain outbreaks, an incorrect in-control parameter is estimated.

When the in-control mean is time varying due to long- and short-term trends (such as seasonality) matters become more complicated and only few approaches exist in the literature. Two methods operating with time-varying parameters for count data are the method by Rossi et al. (1999) and Rogerson and Yamada (2004a). Letting $\mu_{0,t}$ be the time varying in-control mean the first suggests a transformation to normality by looking at

$$x_t = \frac{y_t - 3\mu_{0,t} + 2\sqrt{\mu_{0,t} \cdot y_t}}{2\sqrt{\mu_{0,t}}} \quad (4)$$

and applying a Gaussian CUSUM to these transformed values. However, simulation studies show that in case of low-counts the resulting ARLs are far away from the anticipated ARLs as computed for the Gaussian CUSUM (Rogerson and Yamada, 2004a; Höhle and Paul, 2008). This lead Rogerson and Yamada (2004a) to propose time-varying control parameters of the Poisson CUSUM in order

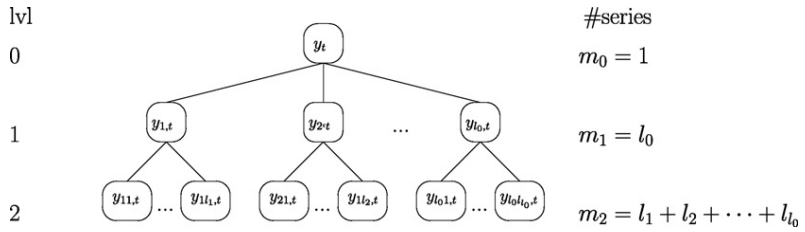


Fig. 1. Hierarchical time series structure. Each time series $y_{X,t}$ is formed by aggregating its immediate descendants in the graph, e.g. $y_t = \sum_{i=1}^{l_0} y_{i,t}$.

to obtain a specific in-control ARL_0 value of γ. Their CUSUM is

$$S_n = \max(0, S_{n-1} + h_n(x_n - k_n)), \quad \text{with}$$
$$k_n = \frac{\mu_{1,n} - \mu_{0,n}}{\log(\mu_{1,n}) - \log(\mu_{0,n})},$$

and $\mu_{1,n}$ being a multiple of standard deviations larger than $\mu_{0,n}$. The factor $h_n = c/c_n$ scales the contribution of $(x_n - k_n)$ at each time point. The threshold c_n is determined at each time point as the threshold of a time-constant Poisson CUSUM with reference value k_n having ARL_0 equal to γ, which can be computed, e.g. by the algorithm of Hawkins (1992). Finally, c is the threshold of an ordinary Poisson CUSUM with constant in-control mean parameter μ_0, e.g. selected as the mean of the training period observations. An alarm is given if $S_n \geq c$.

As an alternative Höhle and Paul (2008) suggest using (2) directly, which makes allowance for quite flexible models for $\mu_{1,t}$ at the cost of loosing the recursive computation in (3) and having to compute performance criterion by Monte-Carlo simulation. Big advantage of the CUSUM methods is their ability for optimal detection of change-points from $\mu_{0,t}$ to $\mu_{1,t}$, however relying on these models being adequate. Care thus has to be exercised in order to specify a reasonable in-control model such as (1). When monitoring massive amounts of time series such care is not always possible. Also, for extremely low counts it can be beneficial to monitor the number of zero count periods before a period with one or more counts by the geometric distribution (Bourke, 1992).

2.2. Multivariate and hierarchical surveillance

Practical statistical surveillance typically means considering multiple time series simultaneously, e.g. series for different diseases and serotypes, age groups or distinct geographical regions. A naive way to perform multivariate monitoring of m time series $\{y_{i,t}; i = 1, \ldots, m\}$ is to apply the univariate method of choice to each time series separately and report all alarms. However, this approach ignores any correlations between the time series and thus leads to inferior detection. Multivariate change-point detection methods such as multivariate CUSUMs take correlations into account, but they are only developed for the continuous case. Rogerson and Yamada (2004b) investigate one of several proposals for a multivariate CUSUM in a surveillance context. However, the more detailed the partition due to serotype, age or region, the rarer the cases. Continuous approximations are thus unsound, but methods for handling correlated multivariate count data time series, e.g. Tourneret et al. (2002), are still scarce. As a consequence, operating with multiple univariate detectors is still the pragmatic choice.

One issue also not dealt with in the literature is the important question of choosing the appropriate level of aggregation for the MOSS time series. Fig. 1 shows a hierarchy of time series resulting from aggregation of three levels, e.g. spatial aggregation with top level being the total aggregated number of cases in a country and lower levels representing the series for administrative regions and districts.

The simple, yet very effective idea is now to monitor all m time series independently by univariate methods. To keep in-control alarm rates comparable between levels one would use higher thresholds at the lower levels. A subsequent plot showing all generated alarms provides a good overview of aggregation effects and outbreak sizes.

2.3. Results

As an example of statistical surveillance in veterinary public health, Fig. 2 shows the 1985–2006 time series of monthly incidence of rabies among foxes in the federal state of Hesse, Germany. These data are part of the monitoring database kept by the Collaboration Centre for Rabies Surveillance and Research (WHO Collaboration Centre for Rabies Surveillance and Research, 2007).

A drastic decrease in the number of cases is seen as a consequence of the oral rabies vaccination program started in 1985 using Tübingen baits. However, several set-backs for the boarder region between Hesse and Bavaria occurred as mentioned in Müller et al. (2005) and shown in the following analyses. To illustrate seasonality of the time series we proceed as in Harnos et al. (2006) and divide the monthly cases by the respective yearly average and compute monthly means of this detrended time series

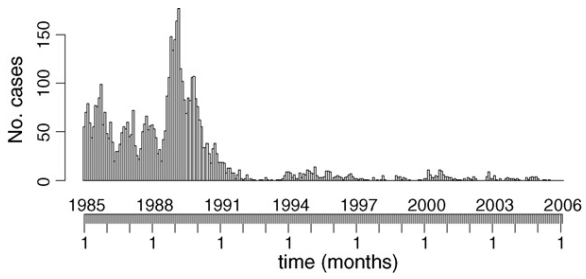


Fig. 2. Monthly number of rabies cases in Hesse, Germany.

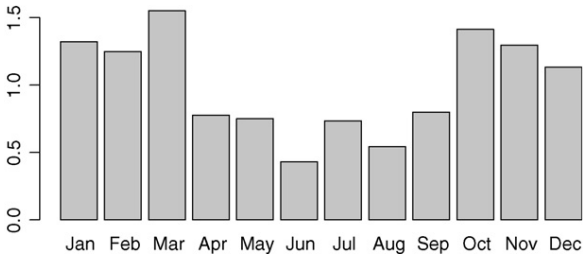


Fig. 3. Seasonality in the rabies cases illustrated by the monthly average relative to yearly average for each month.

as shown in Fig. 3. Strong seasonality of the rabies data can be seen – the increased incidence of up to 1.5 times the yearly average in spring and autumn corresponds to the mating season and dispersal of young foxes (Thulke et al., 2000).

Fig. 4 shows the result of prospective surveillance for the Hesse time series, beginning from January 1998. To obtain the time varying $\mu_{0,t}$ values for each time point t, a seasonal Poisson model as in (1) is fitted to the observed values from January 1985 up to December 1997. As observations close to t are considered more informative, a weighted Poisson regression is performed, where each observation is weighted according to its distance to t in time. This fitted model is then used to compute the predicted in-control mean for all future y_t. Lines show the resulting $\mu_{0,t}$ values and the resulting value of the LR(n) statistic for a 50% increase, i.e. with $\mu_{1,t} = (3/2) \cdot \mu_{0,t}$. With a threshold of $c = 4.0$ the probability for a false alarm within the monitored 98 timepoints is calculated to be 0.04 – with this threshold the first outbreak is detected March 2000. After the alarm the detector is reset by re-estimating the Poisson model now using values up to March 2000 as historical values. Subsequent predictions for the in-control mean are then used to continue monitoring from April 2000 until the next alarm. This procedure continues until the end of the monitoring period is reached – resulting alarms (triangles) and in-control mean, $\mu_{0,t}^{\text{re}}$, obtained after re-fitting following each alarm, are shown in Fig. 4.

As a second step taking spatial aggregation into account we consider the three levels of aggregation arising from the administrative division in Hesse, see Fig. 5. However, due to an administrative reform data on district level are only

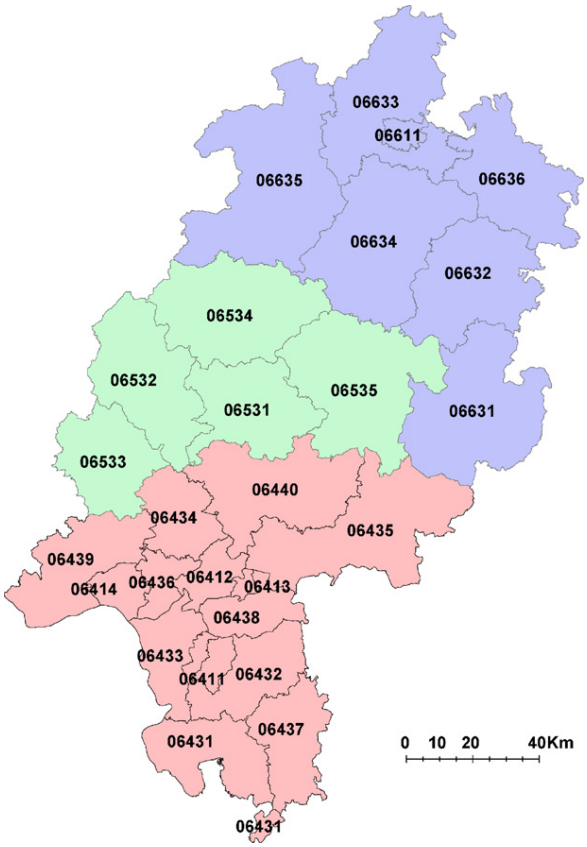


Fig. 5. Map of Hesse, Germany. The shaded regions indicate the three different administrative regions (south, middle, north), the numbers are the unique identifiers (Gemeindeschlüssel) of the 26 (14 + 5 + 7) districts of Hesse.

available from 1990. Fig. 6 shows surveillance of all time series of the hierarchy beginning from January 1998. To avoid careful tuning, the Farrington method with $w = 2$ and $b = 4$ is used.

It is immediately clear that the outbreak during 2000 for the entirety of Hesse was due to problems in the southern region – especially district 06435 (Main-Kinzig-Kreis) located at the border to Bavaria. Investigations showed that there were problems with synchronising

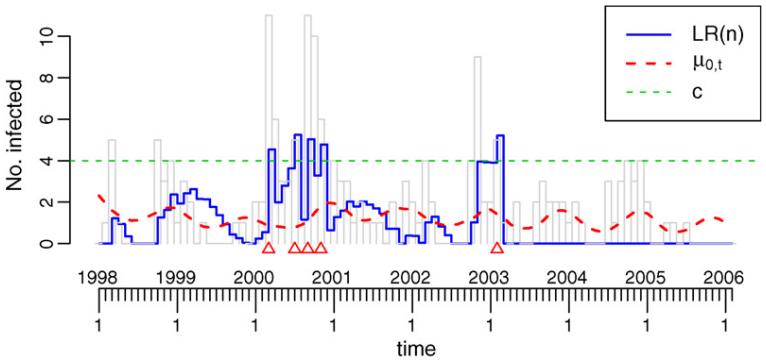


Fig. 4. Statistical surveillance of the Hesse data using the LR-CUSUM method.

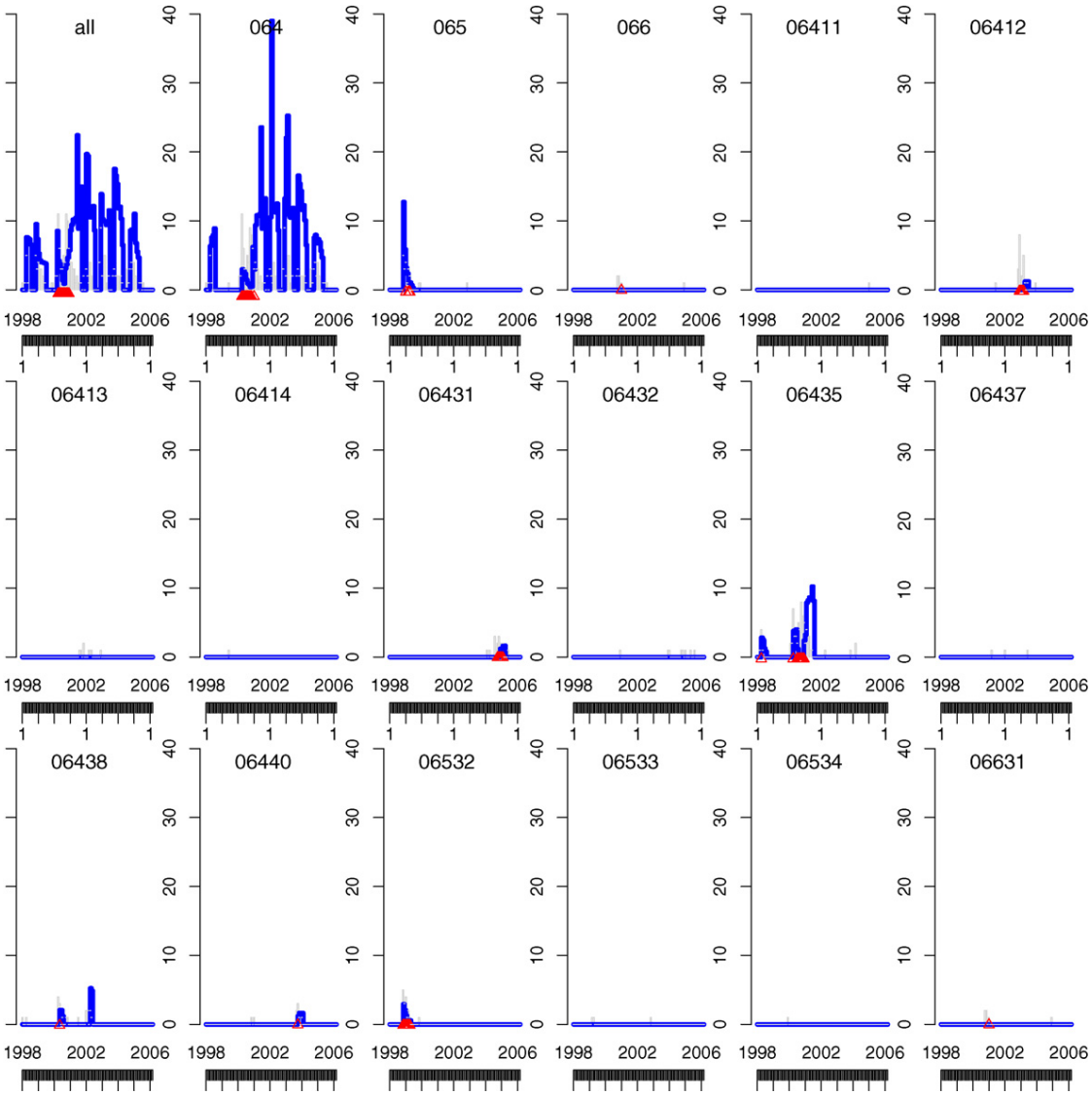


Fig. 6. Surveillance results for each of the $m = 30$ time series having at least one case in the monitored period. Top-left plot is of all cases in Hesse followed by the three administrative regions and the districts. For better visualization the y-axis is truncated at the value 40.

aerial and hand distribution of the oral rabies vaccination in this and the neighbouring city of Offenbach (06413) (Müller et al., 2005). Fig. 7 summarizes the results in a so called alarm plot, which shows a problem in district 06532 (Lahn-Dill-Kreis), causing alarms for region middle at the turn of 1999.

3. Retrospective analysis of surveillance data

The focus of prospective surveillance, described in the previous section, is on outbreak detection. In contrast, retrospective surveillance tries to explain temporal and spatio-temporal patterns in the data through statistical modelling. Following Held et al. (2005), one possible model approach for the analysis of multivariate surveillance data is presented.

3.1. Multivariate modelling

As in Section 2.2, let $y_{i,t}, i = 1,\ldots,m, t = 1,\ldots,n$, denote a multivariate time series of counts. The main feature of the model proposed by Held et al. (2005) is the additive decomposition of the incidence into an endemic and an epidemic component with rates $\eta_{i,t}$ and $\nu_{i,t}$, respectively. In the simplest case, the observed counts $y_{i,t}$ in region i at time t are assumed to be Poisson distributed with mean

$$\mu_{i,t} = \eta_{i,t} + \nu_{i,t}\,.$$

Endemic incidence is persistent with a stable temporal pattern. The endemic component $\nu_{i,t}$ thus may include terms for long-term trends and seasonality and is basically modelled as in (1). Region-specific intercepts allow for

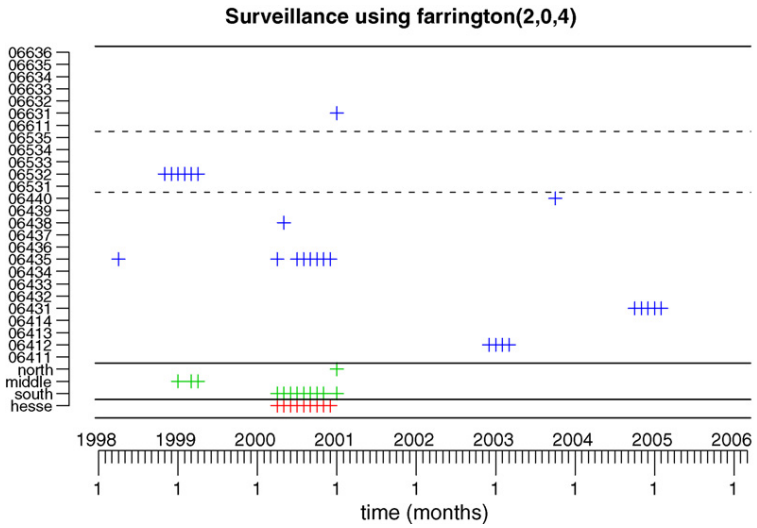


Fig. 7. Alarm plot resulting from the hierarchical surveillance of the $m = 1 + 3 + 26$ time series using the Farrington method. The solid lines divide the series according to their hierarchy (at bottom is lvl. zero), dashed lines group descendants.

different incidence levels in the m regions. All in all, $\nu_{i,t}$ is specified as

$$\log \nu_{i,t} = \alpha_i + \beta t + \sum_{s=1}^{S}\left[\gamma_s \sin\left(\frac{2\pi}{r} s \cdot t\right) + \delta_s \cos\left(\frac{2\pi}{r} s \cdot t\right)\right].$$

The epidemic component $\eta_{i,t}$ should be able to explain occasional outbreaks and capture spatio-temporal dependence caused by the spread of the disease across regions. One possible approach is to let the number of previous counts in the region and in neighbouring regions enter as autoregressive covariates in the epidemic component, i.e.

$$\eta_{i,t} = \lambda y_{i,t-1} + \phi \sum_{j \sim i} y_{j,t-1}$$

where $j \sim i$ denotes all regions adjacent to region i. As all conditioning is on previous values model inference can be performed by maximum likelihood (ML) (Held et al., 2005; Paul et al., 2008).

In many applications the Poisson assumption of equal mean and variance is not realistic. To adjust for possible overdispersion, Held et al. (2005) suggest a negative binomial model with additional dispersion parameter $\psi > 0$, where the mean remains the same but the variance increases to $\mu_{i,t} + \mu_{i,t}^2/\psi$.

3.2. Results

To illustrate the modelling approach described in Section 3.1 we consider the monthly number of rabies cases in Hesse and Bavaria. The data are analysed on two aggregation levels: state level and district level. Data on district level are not available until 1990, hence we only use data for 1990–2006. As described in Section 2.3, there is a strong seasonality and the number of cases is decreasing as a consequence of the vaccination program. Therefore, seasonal terms and a linear time trend are always included.

Separate univariate analyses of the counts in Hesse and in Bavaria showed that incidence levels and the slopes of the linear trend differ whereas seasonality is similar in both federal states. Likelihood-ratio tests suggest to use $S = 2$ seasonal terms. Furthermore, there is evidence for overdispersion because the negative binomial models result in a significant increase in terms of maximised log-likelihood compared to the corresponding Poisson models.

Table 1 shows results for the joint analysis on state level. In all models, $S = 2$ seasonal terms and a state-specific linear trend are included in the endemic component. Inclusion of an autoregressive parameter λ leads to a pronounced increase of the likelihood. The ML estimate of λ is 0.69 (0.05), clearly indicating a temporal dependence after adjustment for seasonal effects. Inclusion of the autoregressive parameter ϕ does improve the fit only slightly and is not required according to the AIC model choice criterion.

On district level, we restrict our attention to 12 districts in the boarder region between Hesse and Bavaria (nine

Table 1
Summary of ML estimates (standard errors) of different negative binomial models for the rabies data in Hesse and Bavaria (state level), $\log L$ denotes the maximised log likelihood, p is the number of parameters and AIC $= -2\log L + 2p$.

$\hat{\lambda}_{ML}$ (se)	$\hat{\phi}_{ML}$ (se)	$\hat{\psi}_{ML}$ (se)	$\log L$	p	AIC
–	–	1.04 (0.12)	−829.1	9	1676.2
0.69 (0.05)	–	4.76 (0.96)	−702.8	10	1425.6
0.68 (0.05)	0.022 (0.019)	4.87 (0.99)	−701.9	11	1425.8

Table 2
Summary of ML estimates (standard errors) of different models for the rabies data in 12 districts in the boarder region of Hesse and Bavaria, $\log L$ denotes the maximised log likelihood, p is the number of parameters and AIC $= -2\log L + 2p$.

$\hat{\lambda}_{ML}$ (se)	$\hat{\phi}_{ML}$ (se)	$\hat{\psi}_{ML}$ (se)	$\log L$	p	AIC
–	–	0.22 (0.02)	−1359.2	16	2750.5
0.57 (0.05)	–	0.82 (0.12)	−1167.8	17	2369.7
0.55 (0.05)	0.041 (0.008)	0.91 (0.14)	−1146.6	18	2329.3

districts in Hesse and three districts in Bavaria). Two districts have been defined to be adjacent if they share a common border. Results for several models on this district level are shown in Table 2. As above, we chose the negative binomial model and included a linear trend and $S = 1$ seasonal terms in the linear predictor, higher terms for seasonality did not lead to a significant improvement in the likelihood. Again, there is evidence for temporal dependence. In addition, there exists spatial dependence: the autoregressive parameter ϕ that captures the influence of neighbouring districts contributes markedly to a better fit. This finding is consistent with the spread of the disease across districts described in Müller et al. (2005).

4. Discussion

Surveillance and monitoring in veterinary public health is an important contribution for the detection and control of diseases in veterinary epidemiology. An important criterion for selecting the appropriate outbreak detection method in the MOS system is the number of time series to monitor. If performance is premium making the cost of tuning second-rank, the Rogerson and Yamada (2004a) or the direct LR-CUSUM method from Section 2.1.2 should be used. When there is uncertainty about the correct parameter of the alternative so called generalized likelihood ratio (GLR) detectors can be used, which estimate the unknown parameter (Höhle and Paul, 2008).

Statistical surveillance of multivariate count time series is an active research area. Our approach of multiple univariate surveillance ignores correlations between the series. With an appropriate time series model explaining seasonality the remaining correlation is often negligible or caused by auto-regression due to disease transmission. One possibility is thus to let the out-of-control model contain an autoregressive component for transmission within and between units as suggested in Section 3.2. One challenge in veterinary applications of such multivariate disease monitoring is the heterogeneity between farms, e.g. when looking at mortality or abortions. If detection is based on per farm basis, the statistical parameter estimation guarantees that monitoring occurs at the level specific to each farm. However, when operating on data of clusters of farms other strategies have to be adopted. Furthermore, we did not discuss the actions taken once the outbreak detection algorithms sound an alarm, because the specific control options depend very much on the objectives of the veterinary MOSS as discussed, e.g. in Christensen (2001).

A retrospective analysis of surveillance data may give clues about temporal and spatio-temporal patterns of the disease considered. In Section 3.2, we were able to identify a significant autoregressive coefficient between neighbouring units, which is one way of quantifying the spatio-temporal behaviour of fox rabies using a time series oriented approach with a persistent spatial pattern of districts or regions. In current work we consider extended model approaches for the analysis of spatio-temporal surveillance data (Paul et al., 2008; Höhle, 2008).

As a consequence of the continuing efforts on rabies surveillance, Germany was as of 1 April 2008 declared rabies free (Freuling et al., 2008). This shows the importance of a structured MOSS for disease eradication and underlines that rabies efforts are worthwhile. In many aspects rabies is an overlooked disease, because whilst in industrial countries it is about to go extinct, it still causes a substantial number of deaths in developing countries as accentuated, e.g. by the 2008 World Rabies Day (Anonymous, 2008). Other examples of monitoring the effect of control measures during outbreak control are the foot and mouth disease outbreak in the UK (Ferguson et al., 2001) or classical swine fever in the Netherlands (Stegeman et al., 1999).

All statistical outbreak detection methods covered in this paper are available in `surveillance` (Höhle, 2007)– a software package for disease outbreak detection using the free software environment for statistical computing and graphics "R" (R Development Core Team, 2008).

Conflict of interest statement

None declared.

Acknowledgments

We thank Christoph Staubach, Federal Research Institute for Animal Health, Germany, for helpful comments and for providing us with the rabies data. Financial support was given by the German Science Foundation (DFG, 2003-2006) and the Swiss National Science Foundation (SNF, since 2007). The first author also thanks the Munich Center of Health Sciences for financial support.

References

Anonymous, 2008. World rabies day. http://www.worldrabiesday.org/.
Bourke, P., 1992. Performance of cumulative sum schemes for monitoring low count-level processes. Metrika 39, 365–384.
Bravata, D.M., McDonald, K.M., Smith, W.M., Rydzak, C., Szeto, H., Buckeridge, D.L., Haberland, C., Owens, D.K., 2004. Systematic review: surveillance systems for early detection of bioterrorism-related diseases. Annals of Internal Medicine 140 (11), 910–922.
Buckeridge, D., Burkom, H., Campbell, M., Hogan, W., Moore, A., 2005. Algorithms for rapid outbreak detection: a research synthesis. Journal of Biomedical Informatics 38 (2), 99–113.
Carpenter, T.E., Chriel, M., Greiner, M., 2007. An analysis of an early-warning system to reduce abortions in dairy cattle in Denmark incorporating both financial and epidemiologic aspects. Preventive Veterinary Medicine 78, 1–11.
Chen, R., 1978. A surveillance system for congenital malformations. Journal of the American Statistical Association 73, 323–327.
Christensen, J., 2001. Epidemiological concepts regarding disease monitoring and surveillance. Acta Veterinaria Scandinavica 94, 11–16.
Diggle, P.J., Rowlingson, B., Su, T.-L., 2005. Point process methodology for on-line spatio-temporal disease surveillance. Environmetrics 16, 423–434.
Doherr, M., Audigé, L., 2001. Monitoring and surveillance for rare health-related events: a review from the veterinary perspective. Philosophical Transactions of the Royal Society B: Biological Sciences 356, 1097–1106.
Farrington, C., Andrews, N., 2003. Outbreak detection: application to infectious disease surveillance. In: Monitoring the Health of Populations. Oxford University Press, (Ch. 8), pp. 203–231.
Farrington, C., Andrews, N., Beale, A., Catchpole, M., 1996. A statistical algorithm for the early detection of outbreaks of infectious disease. Journal of the Royal Statistical Society, Series A 159, 547–563.
Ferguson, N.M., Donnelly, C.A., Anderson, R.M., 2001. Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. Nature 413, 542–548.
Freuling, C., Selhorst, T., Kliemt, A., Conraths, F.J., Müller, T., 2008. Deutschland ist tollwutfrei. Forschungsreport des Bundesministeriums für Ernährung, Landwirtschaft und Verbraucherschutz (1).

Frisén, M., 2003. Statistical surveillance: optimality and methods. International Statistical Review 71 (2), 403–434.
Harnos, A., Reiczigel, J., Solymosi, N., Vattay, G., 2006. Analysis of the effect of immunization in rabies time series. Journal of Theoretical Biology 240, 72–77.
Hawkins, D.M., 1992. Evaluation of average run lengths of cumulative sum charts for an arbitrary data distribution. Communications in Statistics, Simulation and Computation 21 (4), 1001–1020.
Held, L., Höhle, M., Hofmann, M., 2005. A statistical framework for the analysis of multivariate infectious disease surveillance data. Statistical Modelling 5, 187–199.
Höhle, M., 2007. Surveillance: an R package for the monitoring of infectious diseases. Computational Statistics 22 (4), 571–582.
Höhle, M., 2008. Spatio-temporal epidemic modelling using additive-multiplicative intensity models. Tech. Rep., Department of Statistics, University of Munich, Germany, no. 41.
Höhle, M., Paul, M., 2008. Count data regression charts for the monitoring of surveillance time series. Computational Statistics and Data Analysis 52 (9), 4357–4368.
Kleinmann, K.P., Abrams, A.M., 2006. Assessing surveillance using sensitivity, specificity and timeliness. Statistical Methods in Medical Research 15, 445–464.
Kosmider, R., Kelly, L., Evans, S., Gettinby, G., 2006. A statistical system for detecting salmonella outbreaks in British livestock. Epidemiology and Infection 134, 952–960.
Kulldorff, M., 2001. Prospective time periodic geographical disease surveillance using a scan statistic. Journal of the Royal Statistical Society, Series A 164, 61–72.
Lai, T., 1995. Sequential changepoint detection in quality control and dynamical systems. Journal of the Royal Statistical Society, Series B 57, 613–658.
Lawson, A., Kleinman, K. (Eds.), 2005. Spatial and Syndromic Surveillance for Public Health. Wiley.
Lucas, J.M., 1985. Counted data CUSUMs. Technometrics 27 (2), 129–144.
Müller, T., Selhorst, T., Pötsch, C., 2005. Fox rabies in Germany – an update. Eurosurveillance 10, 229–231.
Paul, M., Held, L., Toschke, A.M., 2008. Multivariate modelling of infectious disease surveillance data. Statistics in Medicine 27, 6250–6267.
R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0. http://www.R-project.org.
Rogerson, P., Yamada, I., 2004a. Approaches to syndromic surveillance when data consist of small regional counts. Morbidity and Mortality Weekly Report 53, 79–85.
Rogerson, P., Yamada, I., 2004b. Monitoring change in spatial patterns of disease: comparing univariate and multivariate cumulative sum approaches. Statistics in Medicine 23, 2195–2214.
Rogerson, P.A., 2001. Monitoring point patterns for the development of space-time clusters. Journal of the Royal Statistical Society, Series A 164, 87–96.
Rossi, G., Lampugnani, L., Marchi, M., 1999. An approximate CUSUM procedure for surveillance of health events. Statistics in Medicine 18, 2111–2122.
Salman, M.D. (Ed.), 2003. Animal Disease Surveillance and Survey Systems: Methods and Applications. Blackwell.
Sonesson, C., Bock, D., 2003. A review and discussion of prospective statistical surveillance in public health. Journal of the Royal Statistical Society, Series A 5–12.
Stegeman, A., Elbers, A.R.W., Smak, J., de Jong, M.C.M., 1999. Quantification of the transmission of classical swine fever virus between herds during the 1997–1998 epidemic in the Netherlands. Preventive Veterinary Medicine 42, 219–234.
Steiner, S.H., Cook, R.J., Farewell, V.T., Treasure, T., 2000. Monitoring surgical performance using risk-adjusted cumulative sum charts. Biostatistics 1 (4), 441–452.
Thulke, H.-H., Tischendorf, L., Staubach, C., Selhorst, T., Jeltsch, F., Müller, T., Schlüter, H., Wissel, C., 2000. The spatio-temporal dynamics of a post-vaccination resurgence of rabies in foxes and emergency vaccination planning. Preventive Veterinary Medicine 47, 1–21.
Tourneret, J.-Y., Ferrari, A., Letac, G., 2002. Changepoint detection in multivariate Poisson distributions. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, May 13–17, pp. 1573–1576.
WHO Collaboration Centre for Rabies Surveillance and Research, 2007. WHO Rabies - Bulletin - Europe. http://www.rbe.fli.bund.de/.
Widdowson, M.-A., Bosman, A., van Straten, E., Tinga, M., Chaves, S., van Eerden, L., van Pelt, W., 2003. Automated, laboratory-based system using the internet for disease outbreak detection, the Netherlands. Emerging Infectious Diseases 9 (9), 1046–1052.
Wilkinson, L., 1992. Animal and Disease: An Introduction to the History of Comparative Medicine. Cambridge University Press.
Woodall, W., 2006. The use of control charts in health-care and public-health surveillance. Journal of Quality Technology 38 (2), 89–104.

PAPER II

Multivariate modelling of infectious disease surveillance data

Michaela Paul, Leonhard Held & André M. Toschke

Paper published in *Statistics in Medicine*, 2008, **27**, 6250–6267.

STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:6250–6267
Published online 17 September 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3440

Multivariate modelling of infectious disease surveillance data

M. Paul[1], L. Held[1,*,†] and A. M. Toschke[2]

[1]*Biostatistics Unit, Institute of Social and Preventive Medicine, University of Zurich, Switzerland*
[2]*Department of Public Health Sciences, Division of Health and Social Care Research, King's College London, U.K.*

SUMMARY

This paper describes a model-based approach to analyse multivariate time series data on counts of infectious diseases. It extends a method previously described in the literature to deal with possible dependence between disease counts from different pathogens. In a spatio-temporal context it is proposed to include additional information on global dispersal of the pathogen in the model. Two examples are given: the first describes an analysis of weekly influenza and meningococcal disease counts from Germany. The second gives an analysis of the spatio-temporal spread of influenza in the U.S.A., 1996–2006, using air traffic information. Maximum likelihood estimates in this non-standard model class are obtained using general optimization routines, which are integrated in the R package `surveillance`. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: infectious disease surveillance; multivariate time series of counts; space–time models

1. INTRODUCTION

A major challenge in infectious disease epidemiology remains the analysis of data on notifiable diseases typically collected by national surveillance systems. Time series on counts of infectious diseases often show a regular pattern over time such as long-term trends or seasonality but also occasional outbreaks. This mixture of endemic and epidemic behaviours has to be taken into account when modelling such data. Another characteristic is overdispersion with respect to the usual Poisson assumption. Besides, issues such as under-reporting or reporting delays are quite common.

The probably best-studied stochastic models for the spread of an infectious disease over time are mechanistic models such as the chain-binomial model and related continuous time models such as the SIR model [1, 2]. These models directly describe the infection process of the spread from

*Correspondence to: L. Held, Biostatistics Unit, Institute of Social and Preventive Medicine, University of Zurich, Hirschengraben 84, CH-8001 Zurich, Switzerland.
†E-mail: leonhard.held@ifspm.uzh.ch

Contract/grant sponsor: Swiss National Science Foundation (SNF)

Received 25 March 2008
Accepted 8 August 2008

Copyright © 2008 John Wiley & Sons, Ltd.

person-to-person on an individual level. A key quantity is the basic reproduction number, which together with the number of infectious and susceptibles, allows to provide answers, e.g. about the size and duration of outbreaks or the effect of vaccination programs.

However, a major requirement of epidemic models is that the epidemic process is completely observed. In particular, one has to know at each time point the number of infected and the number of susceptible individuals. If information about susceptibles is not available, the chain-binomial and SIR models can be approximated by a branching process where one assumes an unlimited amount of susceptibles [2]. In large populations such an approximation is especially good since the depletion of susceptibles in the population is, at least at initial stages of an outbreak, negligible. Provided that the proportion of infected individuals stays small relative to the number of susceptibles, the branching process approximation will continue to hold throughout the course of an outbreak [3, 4].

In a surveillance setting, the available data are often spatially and temporally aggregated and information about susceptibles is rarely available. For example, Finkenstädt *et al.* [5, 6] reconstruct the number of susceptibles using data on cases and on births in order to explain extinction and recurrence of epidemics observed in measles. In general, pure mechanistic modelling is too ambitious for routinely collected surveillance data and alternative approaches have been developed.

Purely empirical models, such as log-linear Poisson regression, are not able to capture epidemic outbreaks adequately, so several extensions of Poisson regression models have been suggested. For instance, Zeger and Qaqish [7] propose a Markov regression model with a multiplicative effect of past observations on the disease incidence. For multivariate time series, Knorr-Held and Richardson [8] describe a model where previous counts also enter multiplicatively, modulated by latent binary indicators, which are assumed to follow a two-stage hidden Markov model.

In contrast, Held *et al.* [9] proposed a model based on a branching process with immigration and Poisson offspring. By construction, previous counts now enter additively rather than multiplicatively. The model is extended to allow for seasonality and overdispersion and can be estimated with the maximum likelihood (ML) techniques. This approach yielded promising results in the analysis of univariate and multivariate time series on a specific disease caused by a single pathogen.

However, interdependencies between diseases caused by different pathogens might particularly be of interest to further understand the dynamics of such diseases. For example, viral infections may cause physical damage to respiratory cells and thus facilitate bacterial adherence, clearing the way for bacterial disease [10]. Several studies give both clinical and epidemiological support that influenza infections predispose meningococcal disease [11–14]. For Illustration, Figure 1 shows the weekly number of influenza (labelled as FLU) and meningococcal disease cases (labelled as MEN) in Germany, 2001–2006, obtained from the German national surveillance system for notifiable diseases, administered by the Robert Koch Institute (RKI) [15]. The influenza counts show yearly outbreaks of different severity during the winter. The meningococcal disease counts also display a seasonal pattern with small outbreaks during the winters of 2003 and 2005, which seem to coincide with the two biggest outbreaks of influenza.

Another possible scenario is that several infections are transmitted via the same route. If there is an underlying increase in transmission, e.g. due to varying contact rates, this will induce a directionless correlation between the diseases considered. For example, Farrington *et al.* [16] analyse data on several airborne infections while De Angelis *et al.* [17] consider several sexually transmitted diseases.

In this paper we extend the multivariate model introduced in Held *et al.* [9] to analyse data from different pathogens. In particular, overdispersion as well as seasonality is allowed to vary across diseases and additional parameters capturing a directed influence from one disease to the other are

Copyright © 2008 John Wiley & Sons, Ltd.
Statist. Med. 2008; **27**:6250–6267
DOI: 10.1002/sim

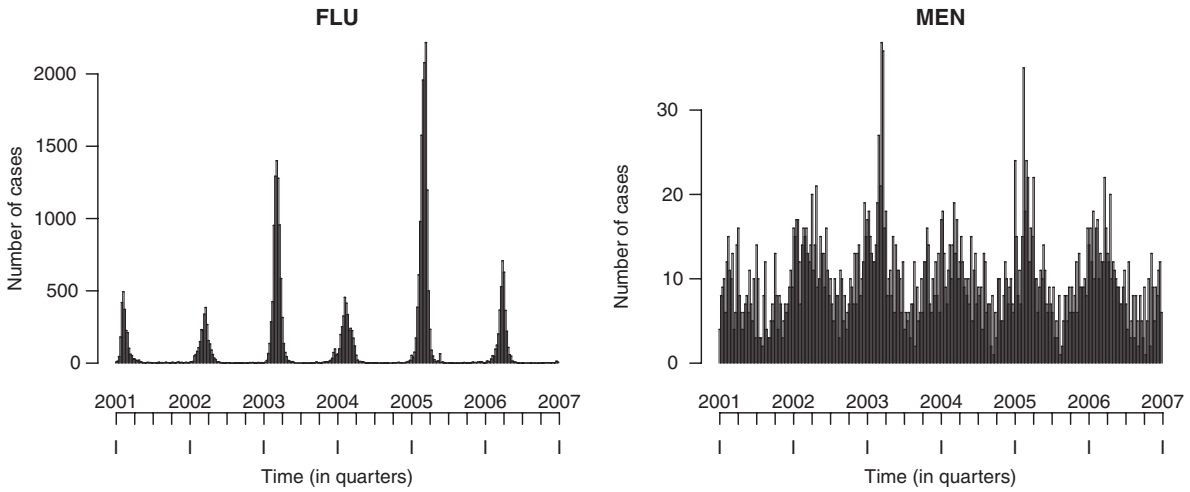


Figure 1. Weekly number of influenza (FLU) and meningococcal disease (MEN) cases in Germany, 01/2001–52/2006.

introduced as described in Section 2. In Section 3.1 we use this model framework to investigate a possible directed association between influenza and meningococcal disease cases for the data described above. Furthermore, this analysis gives empirical evidence of the better fit of additive rather than multiplicative models.

The multivariate branching process model proposed in Held *et al.* [9] can also be applied to spatio-temporal surveillance data. In Section 2.1, we describe a further extension of the model by including additional weights. External data such as information on travel intensities between spatial units can now be incorporated. As an example, in Section 3.2 we consider the weekly number of deaths from influenza and pneumonia in the U.S.A., previously analysed in Brownstein *et al.* [18]. A multivariate model including air traffic information provides a better fit than simple adjacency-based models or models without interdependencies between geographical regions. We close with some discussion in Section 4.

2. THE MODEL FRAMEWORK

To begin, let $y_{i,t}$ denote the number of cases observed in 'unit' i at time t, $i=1,\ldots,m$, $t=1,\ldots,T$. A unit might represent not only a single disease observed, e.g. in several geographical regions or in different age groups, but also different pathogens observed in one location. We might, for example, be interested in multiple diseases transmitted via the same route, e.g. airborne infections. A simple version of the model considered assumes that the counts are negative binomial distributed, $y_{i,t}|y_{i,t-1}\sim\text{NegBin}(\mu_{i,t},\psi)$, with conditional mean

$$\mu_{i,t}=\lambda y_{i,t-1}+\exp(\eta_{i,t}) \tag{1}$$

and conditional variance

$$\mu_{i,t}(1+\psi\mu_{i,t})$$

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

here $\psi>0$ is an additional overdispersion parameter, see below for details. The disease incidence $\mu_{i,t}$ can thus be decomposed additively into two parts. Following Held *et al.* [9] we call the first part $\zeta_{i,t}=\lambda y_{i,t-1}$ (where λ is an unknown autoregressive parameter) the 'epidemic' component, and the second part $\nu_{i,t}=\exp(\eta_{i,t})$ the 'endemic' component. In Section 2.2 we describe parametric approaches to model $\nu_{i,t}$.

The epidemic component should be able to capture occasional outbreaks whereas the endemic component explains a baseline rate of cases that is persistent with a stable temporal pattern. For example, it is quite common to distinguish 'sporadic' and epidemic outbreaks in the incidence analysis of meningitis [8].

The negative binomial model allows for possible overdispersion due to, for example, under-reporting or unobserved covariates that affect the disease incidence. For $\psi=0$ the negative binomial model reduces to a Poisson model where the conditional variance is equal to the conditional mean $\mu_{i,t}$. Note that Held *et al.* [9] use $1/\psi$ as the overdispersion parameter.

In the above model, overdispersion is identical in every unit. If the units are age groups or regions, this may be a realistic assumption. However, when the units correspond to different types of diseases this assumption is unlikely to hold and unit-specific overdispersion parameters ψ_i may be used instead.

2.1. Epidemic component

The epidemic component in (1) is modelled by an autoregression on the number of cases $y_{i,t-1}$ in unit i at the previous time point $t-1$. The inclusion of previous cases allows for temporal dependence beyond seasonal patterns within a unit. However, the model will not be able to explain the spread of a disease across units. Hence, Held *et al.* [9] suggest to include the sum of the previous number of cases $y_{j,t-1}$ in other units $j\neq i$ as a potential explanatory variable for the disease incidence in unit i. Depending on the context, the other units may be e.g. all other units or solely geographically neighbouring units.

Here, we consider a more general version of the epidemic component

$$\zeta_{i,t}=\lambda_i y_{i,t-1}+\phi_i \sum_{j\neq i} w_{ji} y_{j,t-l} \tag{2}$$

where $y_{j,t-l}$ denotes the number of cases observed in unit j at time $t-l$ with lag $l\in\{1,2,\ldots\}$, and w_{ji} are suitably chosen weights. The simplest choice for the weights is $w_{ji}=1$ for all $j\neq i$, more elaborate choices will be discussed in the following.

The additional autoregressive parameters ϕ_i quantify the influence of $y_{j,t-l}$, $j\neq i$, on $y_{i,t}$. Unit-specific autoregressive parameters are useful if the units correspond to different types of diseases. Depending on the duration of the incubation and infectious period it might also be necessary to look at lagged counts $y_{j,t-l}$ with lag $l>1$. For example, consider both influenza and meningococcal infections. There is experimental as well as epidemiological evidence that respiratory viral infections predispose for bacterial disease [10]. Studies showed that patients with severe meningococcal disease were more likely than control subjects to show serological evidence of recent influenza infection [11, 14]. A possible directed association between influenza and subsequent meningococcal disease in Germany will be analysed in Section 3.1.

Suppose surveillance data on the same pathogen are available for several geographic locations $i=1,\ldots,m$. Possible choices for the weights are then, e.g. $w_{ji}=\mathbb{1}(j\sim i)$, where $\mathbb{1}$ is the indicator function and $j\sim i$ denotes all units that are adjacent to i. Then only regions adjacent to region i

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

are taken into account and the sum of counts in adjacent regions enters as an explanatory variable. Such binary weights have been used in Held *et al.* [9] to analyse the weekly number of measles cases in the districts of Lower Saxony, Germany.

However, perhaps more natural is $w_{ji} = \mathbb{1}(j \sim i)/|k \sim j|$, where $|k \sim j|$ denotes the number of neighbours of region j. Thereby we assume that a proportion of infected individuals, say 80 per cent, stays in region j thus being able to infect other individuals in this region. The remaining 20 per cent spread out uniformly among adjacent regions, i.e. individuals in a certain region i (adjacent to j) can also be infected by $20/|k \sim j|$ per cent of cases introduced from region j. This assumption is made for all regions.

The dispersal of cases in space is not necessarily only local but also global. Linking of parallel time series based on adjacencies may then be unrealistic. For instance, SIR-type models on a local level (i.e. cities) have been combined with air transportation data to model the spatio-temporal spread of infectious diseases such as influenza and SARS [19–21]. An alternative choice would be to include travel information in the weights w_{ji} if such information is available. In Section 3.2 we will use the number of airline passengers obtained from the TranStats database, U.S. Department of Transportation [22], to analyse influenza mortality in the U.S.A.

2.2. *Endemic component*

The endemic component includes terms to describe differences between units and seasonality and is specified as

$$\log(\nu_{i,t}) = \eta_{i,t} = \alpha_i + \sum_{s=1}^{S} (\gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t)) \tag{3}$$

where S is the number of harmonics to include and ω_s are Fourier frequencies, e.g. $\omega_s = 2\pi s/52$ for weekly data. Modelling seasonal variation of infectious diseases through superposition of harmonic waves goes back to Serfling [23] and has since then been used in various models (e.g. [13, 24]). An alternative representation of the seasonal terms in (3) as

$$\gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t) = A_s \sin(\omega_s t + \varphi_s)$$

with $A_s = \sqrt{\gamma_s^2 + \delta_s^2}$ and $\tan(\varphi_s) = \delta_s/\gamma_s$ is easier to interpret. Thus, the parameter A_s marks the amplitude of the seasonal component s, while φ_s represents the phase difference [25].

The parameter α_i in (3) allows for different incidence levels in each of the m units (age groups, regions or pathogens). This assumption is reasonable since, e.g. the willingness of a sick person to seek medical advice might depend on the age of that person, for example if case severity does depend on age. Also, the compliance of general practitioners or local public health offices to report cases might differ by area. A further extension of the model is to also let the seasonality terms vary across units, i.e.

$$\log(\nu_{i,t}) = \eta_{i,t} = \alpha_i + \sum_{s=1}^{S_i} (\gamma_{i,s} \sin(\omega_s t) + \delta_{i,s} \cos(\omega_s t)) \tag{4}$$

Note that the number of harmonic waves S_i might also depend on unit i. If we are interested in modelling related diseases, pathogen-specific seasonality is a sensible assumption.

An additional modification is to consider different (standardized) population sizes $n_{i,t}$ by assuming $\mu_{i,t} = \zeta_{i,t} + n_{i,t}\nu_{i,t}$. Time-dependent population adjustment $n_{i,t}$ might be useful for

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

subgroups (age, geographical region, etc.) with a varying distribution over time. However, relatively long time series (covering several years up to decades, say) are typically needed for a pronounced change in population proportions. Adjustment for different but constant proportions among units, n_i, is not necessary since this will be completely absorbed by the unit-specific incidence levels α_i.

2.3. Likelihood inference

Conditional on $y_{i,t-1}, \ldots, y_{i,t-l}$, $i = 1, \ldots, m$, the counts $y_{i,t}$ are assumed to be negative binomial distributed with mean

$$\mu_{i,t}(\boldsymbol{\theta}_i) = \mu_{i,t} = \zeta_{i,t} + \nu_{i,t} \tag{5}$$

where $\zeta_{i,t}$ is given in (2), $\nu_{i,t}$ is given in (4) and $\boldsymbol{\theta}_i = (\lambda_i, \phi_i, \alpha_i, \gamma_{i,1}, \ldots, \gamma_{i,S_i}, \delta_{i,1}, \ldots, \delta_{i,S_i})^{\mathrm{T}}$. Those parameters that are not specified are omitted. With $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m, \psi_1, \ldots, \psi_m)^{\mathrm{T}}$ the log-likelihood is given as

$$l(\boldsymbol{\theta}) = \sum_{i,t} l_{i,t}(\boldsymbol{\theta}_i, \psi_i)$$

where

$$l_{i,t}(\boldsymbol{\theta}_i, \psi_i) \propto \log\Gamma\left(y_{i,t} + \frac{1}{\psi_i}\right) - \log\Gamma\left(\frac{1}{\psi_i}\right) + \frac{1}{\psi_i}\log\left(\frac{1}{1+\psi_i\mu_{i,t}(\boldsymbol{\theta}_i)}\right) + y_{i,t}\log\left(\frac{\psi_i\mu_{i,t}(\boldsymbol{\theta}_i)}{1+\psi_i\mu_{i,t}(\boldsymbol{\theta}_i)}\right)$$

is the log-likelihood contribution of observation $y_{i,t}$ and $\Gamma(\cdot)$ is the gamma function [26, p. 255].

If $\lambda = 0$ and for fixed ψ, model (1) corresponds to a log-linear generalized linear model (GLM) and can be fitted with standard software for GLMs to obtain the ML estimates $\hat{\boldsymbol{\theta}}$ [27]. Otherwise, the log-likelihood $l(\boldsymbol{\theta})$ needs to be optimized numerically using generic optimization routines such as the quasi-Newton method BFGS implemented in the R function `optim`. Held *et al.* [9] used numerical approximations of the score function and Fisher information matrix to fit the model, whereas we use analytical derivatives to speed up the computation of the ML estimates. The fitting procedure is implemented in the R package `surveillance` [28] available from the Comprehensive R Archive Network at `http://cran.r-project.org`. Implementation details can be found in Appendix A.

3. APPLICATION TO DATA

3.1. Influenza and meningococcal disease

Associations between influenza and subsequent rates and severities of meningococcal disease have been well documented in the literature (see e.g. References [11–14] or the reviews of Hament *et al.* [10], Brundage [29] and references therein). We analysed the influence of influenza on invasive meningococcal disease based on the weekly number of cases of both disease types in Germany, 2001–2006 as discussed in Section 1 and shown in Figure 1.

Copyright © 2008 John Wiley & Sons, Ltd.
Statist. Med. 2008; **27**:6250–6267
DOI: 10.1002/sim

Table I. Univariate analysis of meningococcal infections in Germany, 01/2001–52/2006.

S	$\hat{\lambda}$ (s.e.)	$\hat{\psi}$ (s.e.)	$\log L$	p	AIC
1	—	0 (fixed)	−872.1	3	1750.2
1	—	0.06 (0.01)	−850.2	4	1708.3
1	0.16 (0.06)	0.05 (0.01)	−845.6	5	1701.2
2	0.16 (0.06)	0.05 (0.01)	−845.5	7	1705.0

The log-likelihood is denoted by $\log L$, p is the number of parameters and AIC $=-2\log L+2p$.

Note that the influenza data suffer from under-reporting: the observed counts comprise only laboratory confirmed cases and thus, only a fraction of all influenza cases is actually recorded since the illness caused by influenza is often too slight to warrant medical attention. Nevertheless, the data are still able to reflect the temporal course of the disease. On the other hand, a capture–recapture-analysis showed that the degree of ascertainment for meningococcal disease (reported to the RKI) is quite high [30].

Table I summarizes results from a univariate analysis of the meningococcal disease data alone. Based on $S=1$ seasonal term, a negative binomial model instead of the Poisson model with fixed $\psi=0$ results in a significant increase of the maximized log-likelihood from -872.1 to -850.2. Additional inclusion of the autoregressive parameter λ leads to a further substantial improvement of the maximized log-likelihood. The ML estimate of λ is 0.16 (0.06) indicating a weak dependence on the number of cases in the previous week after adjustments for seasonal effects. Higher degrees S for seasonality give only slight improvements of the maximized log-likelihood, so the best model according to the model choice criterion AIC is the negative binomial model with $S=1$ seasonal term and the autoregressive parameter λ, see Table I.

Based on a similar univariate analysis of the influenza data, the best model according to AIC includes the overdispersion parameter, an autoregressive term and $S=3$ seasonal terms.

Table II now summarizes the results of selected multivariate models. The conditional mean of the most general formulation is specified as

$$\begin{pmatrix}\mu_{\text{men},t}\\ \mu_{\text{flu},t}\end{pmatrix}=\begin{pmatrix}\lambda_{\text{men}} & \phi_{\text{men}}\\ \phi_{\text{flu}} & \lambda_{\text{flu}}\end{pmatrix}\begin{pmatrix}\text{MEN}_{t-1}\\ \text{FLU}_{t-1}\end{pmatrix}+\begin{pmatrix}\nu_{\text{men},t}\\ \nu_{\text{flu},t}\end{pmatrix}$$

The first model shown does not allow for any interdependencies between influenza and meningococcal disease (i.e. $\phi_{\text{men}}=\phi_{\text{flu}}=0$) and is based on the respective best univariate models with $\log L$ being the sum of the log-likelihood values from the univariate analyses. The AIC for this model is 3807.5. The second model, which includes an influence from influenza on meningococcal disease (denoted by 'flu → men' in Table II), shows a better fit than the model without interaction (AIC = 3791.9). We also looked at the 'reverse' model that includes an influence from meningococcal disease on influenza ('men → flu'). Since the inclusion of past meningococcal disease cases plays no role, $\hat{\phi}_{\text{flu}}\approx 0$, this supports that the association between meningococcal disease and influenza is directed. Finally, we also fitted a model that includes both 'flu → men' and 'men → flu'. Results for this model are virtually identical to those obtained based on the second model without the additional 'men → flu' term. Indeed, the additional parameter does not improve the log-likelihood $\log L$, therefore AIC is increased by 2 to AIC = 3793.9.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Table II. Multivariate analysis of influenza and meningococcal disease in Germany, 01/2001–52/2006.

$\hat{\lambda}$ (s.e.)		$\hat{\phi}$ (s.e.)		$\hat{\psi}$ (s.e.)			
flu	men	men → flu	flu → men	flu	men	log L	p
0.74 (0.05)	0.16 (0.06)	—	—	0.29 (0.04)	0.05 (0.01)	−1889.7	14
0.74 (0.05)	0.10 (0.06)	—	0.005 (0.001)	0.29 (0.04)	0.04 (0.01)	−1881.0	15
0.74 (0.05)	0.16 (0.06)	4e−07 (1e−04)	—	0.29 (0.04)	0.05 (0.01)	−1889.7	15
0.74 (0.05)	0.10 (0.06)	4e−07 (1e−04)	0.005 (0.001)	0.29 (0.04)	0.04 (0.01)	−1881.0	16

The endemic component $\nu_{i,t}$ includes $S=3$ seasonal terms for the FLU data and $S=1$ seasonal term for the MEN data. The log-likelihood is denoted by $\log L$ and p is the number of parameters.

Figure 2 shows the observed influenza and meningococcal disease cases together with the fitted means $\hat{\mu}_{\text{flu},t}$ and $\hat{\mu}_{\text{men},t}$ for the model with interaction 'flu → men'. The means are separated into two, respectively, three additive components: an endemic and an autoregressive component for the FLU data and an endemic, an autoregressive and an 'influenza-driven' component for the MEN data. The two outbreaks of meningococcal disease in 2003 and 2005 are explained primarily by the influence of influenza, which indicates an association between influenza cases and subsequent meningococcal disease cases. Overall, the model gives a quite good fit to the data and is able to explain the seasonality and the outbreaks in the meningococcal disease data, as well as the different severity of the influenza outbreaks.

Deviance residuals (see e.g. [31]) and corresponding autocorrelation functions are also shown in Figure 2. The residuals seem to be roughly uncorrelated, perhaps with the exception of the lag two autocorrelation for influenza. We also looked at the cumulative periodogram of the residual series to assess compatibility with white noise: no obvious deviation from white noise could be seen. Such tests for white noise, however, neglect the effects of parameter estimation and are therefore only indications and not strictly valid [32].

So far, only the number of influenza cases in week $t-1$ (lag $l=1$) has been considered as explanatory variable for the meningococcal disease incidence in week t. To further investigate the relationship between influenza and meningococcal disease we also considered a delayed effect of influenza towards meningococcal disease of up to three weeks. Table III lists the ML estimates of ϕ_{men} for several negative binomial models with conditional mean for the meningococcal disease data specified as

$$\mu_{\text{men},t} = \lambda_{\text{men}}\text{MEN}_{t-1} + \phi_{\text{men}}\text{FLU}_{t-l} + \nu_{\text{men},t}$$

where the endemic component $\nu_{\text{men},t}$ includes $S=1$ seasonal term, and lags $l \in \{-3, \ldots, 3\}$. Forward lags were also included in the analysis to assess whether the association is unidirectional in time (compare with the analysis in [12]). If there is a directed association the estimated values for ϕ_{men} should be asymmetric around lag zero. As in Hubert *et al.* [12] there is still a positive association for negative lags, but the estimated effect is largest for lag one and slightly smaller for lag zero. This indicates that changes in the incidence rate of influenza might be associated with one-week delayed ($l=1$) or with simultaneous ($l=0$) changes in the incidence rate of meningococcal disease although the evidence is not very strong. Since we only know the week when a case was reported and do not have the exact dates of infection, some event times might be misclassified, which would weaken the autocorrelation and disturb the lag.

For comparison, we also fitted a 'multiplicative' autoregressive conditional model as suggested in Zeger and Qaqish [7] to the meningococcal disease data. In their model, the deviation of the

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

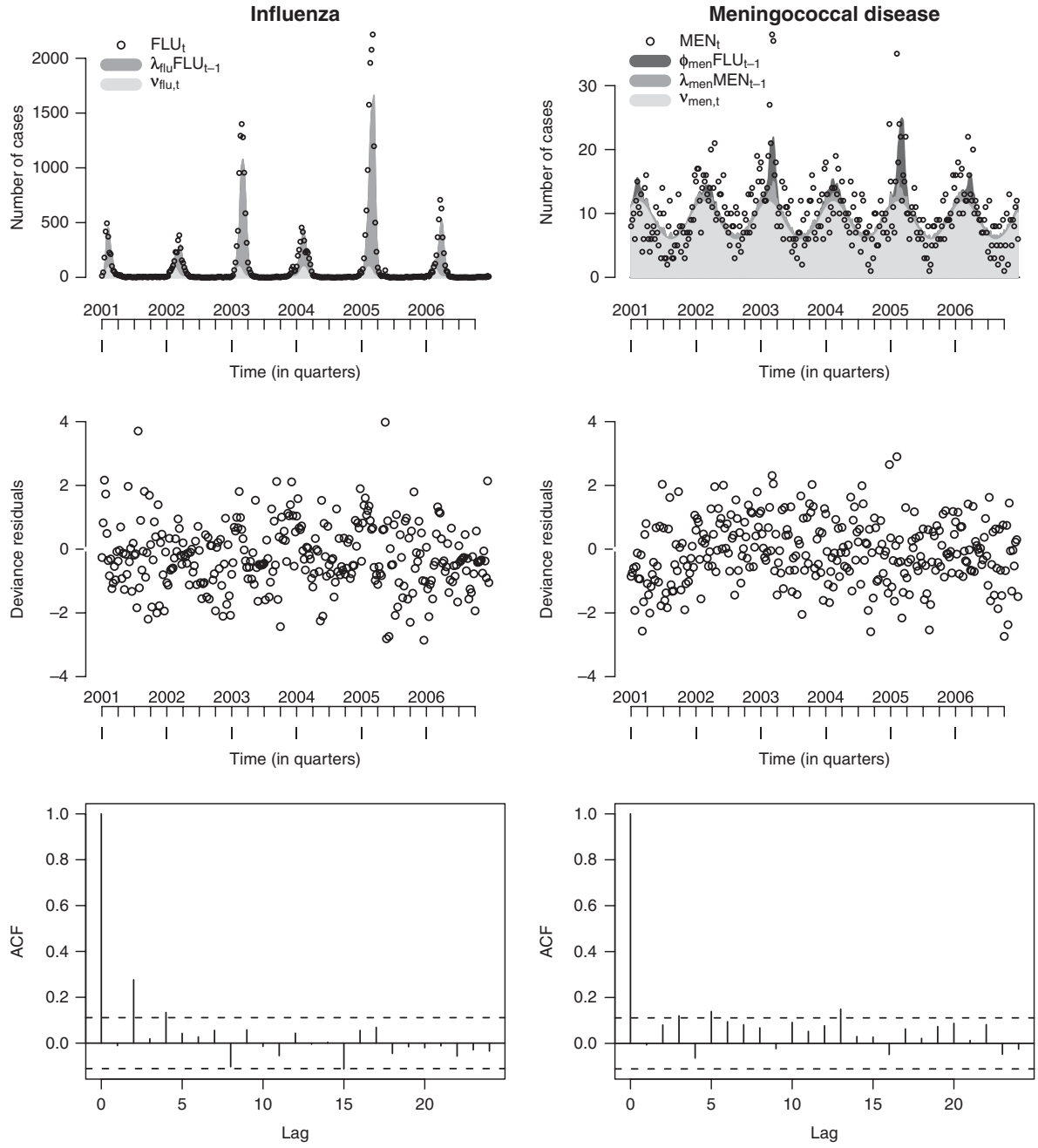


Figure 2. Observed number of influenza and meningococcal disease cases with fitted mean, deviance residuals and corresponding autocorrelation function of the model with interaction 'flu → men'.

logarithm of the observed counts y_{t-1} at time $t-1$ from the linear predictor η_{t-1} at time $t-1$ enters as an explanatory variable:

$$\log(\mu_t) = \eta_t + \theta(\log(y^*_{t-1}) - \eta_{t-1})$$

Copyright © 2008 John Wiley & Sons, Ltd.
Statist. Med. 2008; **27**:6250–6267
DOI: 10.1002/sim

Table III. Analysis of meningococcal disease data with several lagged influenza counts as explanatory variables.

lag l	$\hat{\phi}\times10^3$ (s.e. $\times10^3$)
3	2.92 (1.30)
2	4.54 (1.41)
1	5.32 (1.42)
0	5.30 (1.39)
−1	4.68 (1.31)
−2	3.73 (1.26)
−3	2.30 (1.22)

The mean is specified as $\mu_{\text{men},t}=\lambda_{\text{men}}\text{MEN}_{t-1}+\phi_{\text{men}}\text{FLU}_{t-l}+\nu_{\text{men},t}$, where the endemic component $\nu_{\text{men},t}$ includes $S=1$ seasonal term.

To avoid non-existence of the logarithm, any zero values of y_{t-1} are replaced by a constant c, $0<c<1$, i.e. $y^*_{t-1}=\max\{y_{t-1},c\}$. Note that past cases y_{t-1} are not simply added directly to the linear predictor η_t as would seem natural, because such a model cannot describe positive association without growing exponentially in time [33, Section 10.4]. The conditional mean μ_t of $y_t|y_{t-1}$ is thus given by

$$\mu_t=\exp(\eta_t)\left[\frac{y^*_{t-1}}{\exp(\eta_{t-1})}\right]^\theta \tag{6}$$

so past counts y^*_{t-1} act multiplicatively on the conditional mean μ_t relative to $\exp(\eta_{t-1})$. Extensions and alternative forms of this autoregressive conditional model have been suggested elsewhere [34, 35].

The exact form of the conditional mean μ_t for the meningococcal disease data can be found in Table IV. Throughout, a negative binomial observation model is used and the linear predictor η_t always contains $S=1$ seasonal term. For the multiplicative model (6) any zero values of MEN_t or FLU_t are replaced with $c=0.1$. All AIC values given in Table IV are computed using the log-likelihood contributions of only the MEN data. Letting previous counts of both meningococcal disease and influenza act multiplicatively instead of additively on the mean μ_t gives a worse fit according to AIC. Besides, the introduction of an arbitrary constant c to avoid non-existence of the logarithm is not necessary for the additive model and the parameters are easier to interpret.

3.2. Influenza in the U.S.A.

As part of its national influenza surveillance effort, the Centers for Disease Control and Prevention (CDC) receive weekly mortality reports from 122 cities and metropolitan areas in the U.S.A. within 2–3 weeks from the date of death. These reports summarize the total number of deaths occurring in these cities/areas each week, as well as the number due to pneumonia and influenza. Figure 3 shows the weekly number of deaths from influenza and pneumonia obtained from the CDC 121 Cities Mortality Reporting System for weeks 40/1996 to 39/2006 in nine major geographic regions of the U.S.A. [36]. A map of these regions is shown in Figure 4.

These data have been analysed in Brownstein *et al.* [18]. The authors studied the influence of long-range airline travel on the inter-regional influenza spread and found empirical evidence that

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Table IV. Comparison of 'additive' and 'multiplicative' models for the meningococcal disease data.

Additive model	AIC	Multiplicative model	AIC
$\mu_t = \exp(\eta_t) + \lambda \text{MEN}_{t-1}$	1701.2	$\mu_t = \exp(\eta_t) \left[\frac{\text{MEN}^*_{t-1}}{\exp(\eta_{t-1})} \right]^{\theta_1}$	1703.3
$\mu_t = \exp(\eta_t) + \lambda \text{MEN}_{t-1} + \phi \text{FLU}_{t-1}$	1685.7	$\mu_t = \exp(\eta_t + \theta_2 \log(\text{FLU}^*_{t-1})) \left[\frac{\text{MEN}^*_{t-1}}{\exp(\tilde{\eta}_{t-1})} \right]^{\theta_1}$	1700.4
$\mu_t = \exp(\eta_t) + \phi \text{FLU}_{t-1}$	1686.5	$\mu_t = \exp(\eta_t + \theta_2 \log(\text{FLU}^*_{t-1}))$	1704.9

The endemic component $\nu_t = \exp(\eta_t)$ includes $S=1$ seasonal term, $\text{MEN}^*_t = \max\{\text{MEN}_t, 0.1\}$, $\text{FLU}^*_t = \max\{\text{FLU}_t, 0.1\}$ and $\tilde{\eta}_{t-1} = \eta_{t-1} + \theta_2 \log(\text{FLU}^*_{t-2})$.

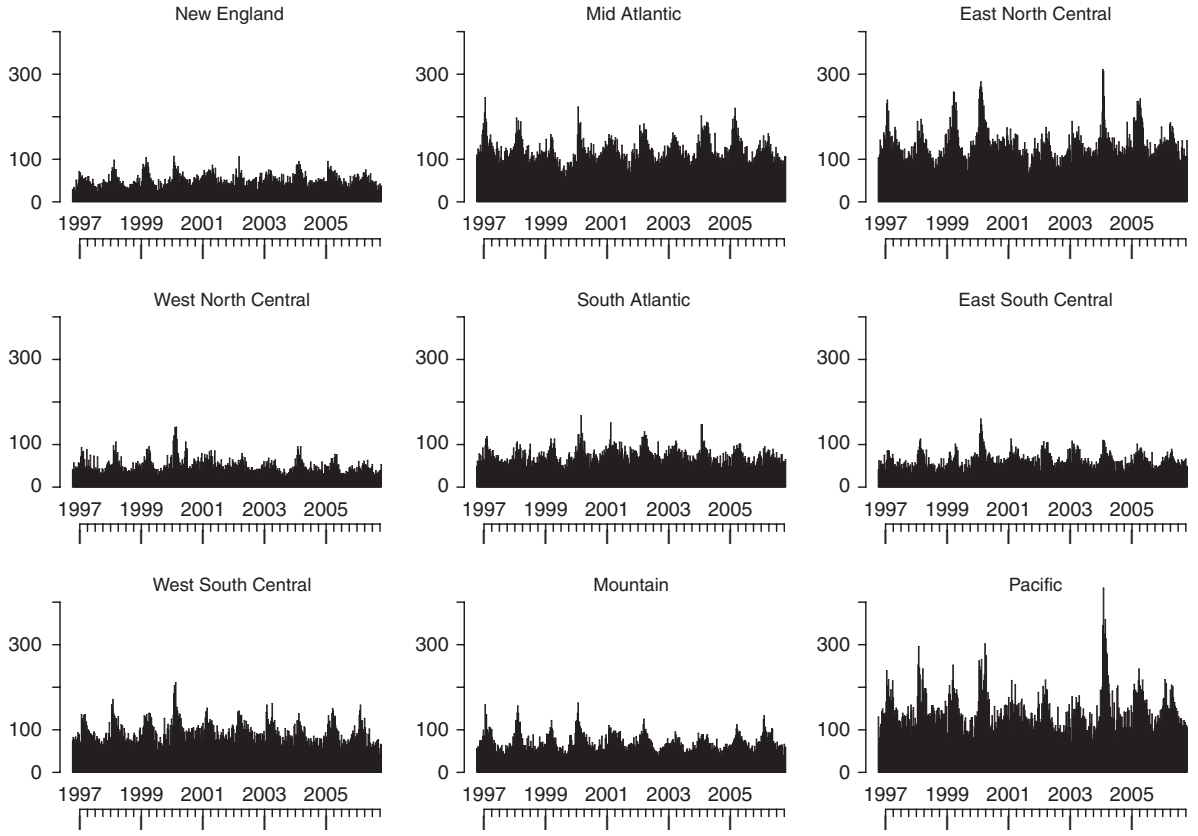


Figure 3. Weekly number of deaths from influenza and pneumonia in the U.S.A. 40/1996–39/2006.

air travel plays a role in the annual spread of influenza in the U.S.A. We applied our model to these data and compared several weights w_{ji} in the epidemic component: weights based on geographic adjacencies and weights including air travel information.

With regard to the form of seasonality $S=4$ seasonal terms are included in the endemic component $\nu_{i,t}$. The epidemic components includes an autoregressive parameter λ as well as region-specific parameters ϕ_i, $i=1,\ldots,9$. As weights in (2) we use firstly $w_{ji} = \mathbb{1}(j \sim i)$, i.e. only neighbouring

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Figure 4. Map of nine major geographical regions of the U.S.A. as defined by the CDC.

regions are considered. Two regions are defined to be neighbours if they share a common border. Secondly, we use geographic weights $\mathbb{1}(j \sim i)/|k \sim j|$ as discussed in Section 2.1. Finally, we use average numbers of passengers travelling by air from region j to region i relative to the population in region j. Data on the yearly number of passengers travelling by air within the U.S.A. were obtained from the TranStats database, U.S. Department of Transportation [22] and data on state population estimates were obtained from the Population Estimates Program, U.S. Bureau of the Census [37]. Denote n_i the population in region i (in the year 2000) and

$$p_{ji} = \#\text{ passengers per week from } j \to i \quad \text{(average over years 1996–2006)}$$

$$p_{ji}(\text{yearly}) = \#\text{ passengers per week from } j \to i \quad \text{(average per year)}$$

That means that the weights p_{ji}/n_j are constant for all time points $t = 1, \ldots, 522$, whereas the weights $p_{ji}(\text{yearly})/n_j$ change each year.

Results of these models are summarized in Table V. Note that only the smallest and largest value of the $\hat{\phi}_i$'s are given in the table instead of all nine values. It can be seen that the inclusion of both the autoregressive parameters λ and ϕ_i improves the fit according to AIC. There is not much difference in terms of maximized log-likelihood between the two models that use geographic weights, the second model that includes the number of neighbours $|k \sim j|$ performing slightly better. Both models that include travel information perform better than the models with geographic weights, the model using time-varying weights $p_{ji}(\text{yearly})/n_j$ being the best.

The actual value of ϕ_i depends on the magnitude of $\sum_j w_{ji} y_{j,t-1}$, therefore a direct comparison of these values from different models is not useful. However, one can look at another quantity. The general model can also be written in a multivariate fashion as

$$\boldsymbol{\mu}_t = \boldsymbol{\Lambda} \mathbf{y}_{t-1} + \boldsymbol{\nu}_t \tag{7}$$

where $\boldsymbol{\mu}_t$, $\mathbf{y}_{t-1}$ and $\boldsymbol{\nu}_t$ are vectors of length m and $\boldsymbol{\Lambda}$ is a $m \times m$ matrix with elements λ_i on the diagonal and elements $(\Lambda)_{ij} = \phi_i w_{ji}$ for $i \neq j$. For constant $\boldsymbol{\nu}_t$ model (7) corresponds to a multivariate branching process with immigration. If the largest eigenvalue of $\boldsymbol{\Lambda}$ is smaller than

Copyright © 2008 John Wiley & Sons, Ltd.
Statist. Med. 2008; **27**:6250–6267
DOI: 10.1002/sim

Table V. Multivariate analysis of influenza mortality in the U.S.A.

w_{ji}	$\hat{\lambda}$ (s.e.)	$\hat{\phi}_i$ (s.e.)	$\hat{\psi}$ (s.e.)	AIC	max EV
—	—	—	0.04 (0.001)	40300.5	—
—	0.34 (0.01)	—	0.03 (0.001)	39693.6	0.34
$\mathbb{1}(j \sim i)$	0.30 (0.01)	0.01 (0.01)–0.23 (0.08)	0.03 (0.001)	39632.2	0.45
$\frac{1}{\lvert k \sim j \rvert} \cdot \mathbb{1}(j \sim i)$	0.30 (0.01)	0.01 (0.02)–0.68 (0.25)	0.03 (0.001)	39631.6	0.44
p_{ji}/n_j	0.28 (0.01)	0.89 (3.13)–31.58 (6.04)	0.03 (0.001)	39617.0	0.45
$p_{ji}(\text{yearly})/n_j$	0.28 (0.01)	0.84 (1.09)–28.68 (5.02)	0.03 (0.001)	39593.5	$*$

The endemic component $\nu_{i,t}$ includes $S=4$ seasonal terms. AIC $=-2\log L+2p$ and max EV denotes the maximum eigenvalue of the estimated matrix $\hat{\mathbf{\Lambda}}$ in (7). The maximum eigenvalues of the model with time-varying weights $(*)$ are shown in Figure 5.

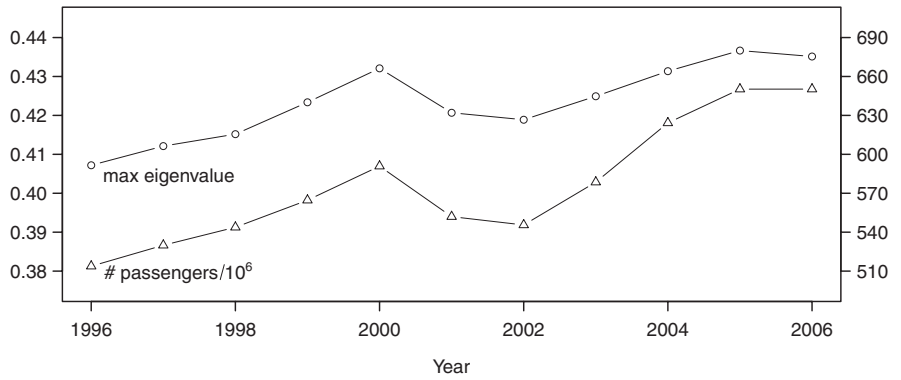


Figure 5. Maximum eigenvalues (circles) of $\hat{\mathbf{\Lambda}}$ for the model with time-varying weights $p_{ji}(\text{yearly})/n_j$ summarized in Table V (left axis). The triangles show the yearly number of passengers per 10^6 (right axis).

unity, the process is ergodic [38]. The largest eigenvalues of $\hat{\mathbf{\Lambda}}$ in the models considered are shown in Table V. Taking spatial variation into account by additional inclusion of $\phi_i \sum w_{ji} y_{j,t-1}$ in the epidemic component leads to an increased maximum eigenvalue compared with the model, which only includes an autoregressive parameter λ. For the model with time-varying weights, the matrix $\hat{\mathbf{\Lambda}}$ changes each year. The respective maximum eigenvalues are shown in Figure 5 together with the yearly number of passengers $\sum p_{ji}(\text{yearly})/10^6$. The development of the number of passengers is mirrored in the curve of the largest eigenvalues.

4. DISCUSSION

In this paper we have proposed a flexible class of statistical models for the analysis of multivariate time series of infectious disease counts. All analyses were done using standard optimization routines where results are readily available in contrast to computer-intensive methods based on Markov chain Monte Carlo (MCMC). A main feature of the model is the decomposition of the disease incidence into an endemic and an epidemic component, which allows to capture occasional

Copyright © 2008 John Wiley & Sons, Ltd.
Statist. Med. 2008; **27**:6250–6267
DOI: 10.1002/sim

outbreaks in the data. The motivation of the epidemic component comes from a branching process formulation well known in infectious disease epidemiology. Note that the interpretation of the branching process as an approximation is only appropriate if the generation time, i.e. the time between 'generations' of infectives, equals the observation time at which data are collected (i.e. days, weeks or months). However, simulation studies showed that a Poisson branching process, aggregated to coarser time intervals, can be approximated by a branching process with additional overdispersion [9].

In our example from Section 3.1 the assumption of disease-specific parameters proved to be reasonable. There are much more influenza cases than meningococcal disease cases and therefore the amplitudes of the seasonal patterns differ a lot. Besides, the influenza and meningococcal disease data showed a different amount of overdispersion. The use of weights w_{ji} in the epidemic component may lead to a better accounting for spatial dependence. For instance, the analysis of influenza and pneumonia mortality in nine major regions of the U.S.A. in Section 3.2 showed that the inclusion of air travel information yielded a better model in terms of AIC.

A further model generalization, which we are currently considering, is the introduction of random effects in the endemic or epidemic component (e.g. for incidence levels or for the autoregressive parameters). If the number of units m is large, the number of unit-specific parameters may otherwise get large rather quickly, which might lead to identifiability problems. Estimates of parameters can then be obtained by optimizing the marginal likelihood, which involves integration of the likelihood with respect to the distribution of the random effects. This integral cannot be computed explicitly and must be approximated, e.g. using adaptive Gaussian quadrature or the Laplace approximation for the integrand.

Seasonality is modelled with superposition of sine and cosine terms in the endemic component, which implies that the phase and the amplitude of the seasonal pattern is constant across all considered periods. This might not always be adequate, e.g. if there is a year-by-year variation in the starting point of the seasonal pattern. The formulation is, however, not restricted to this specific modelling of seasonal variation, other choices would also be possible. For example, Fanshawe *et al.* [39] use harmonic regression to model seasonal variations in particulate matter concentrations. To deal with yearly variation of the seasonal pattern the static parameters γ and δ of the sine and cosine terms are replaced with independent random walks.

In this paper a retrospective analysis of time series of counts of infectious diseases was of main interest. The comparison of the different models was based on the model choice criterion AIC. An alternative is to use BIC, in particular if the quality of one-step-ahead predictions is of interest [40]. If the prediction of future observations is a main goal, the validity of models can also be assessed with proper scoring rules that evaluate a model based on the prediction and the actual observed value [41, 42].

Another extension not considered here is the introduction of time-varying parameters, in particular time-varying autoregressive parameters λ_t or ϕ_t as suggested by Held *et al.* [25] where the autoregressive parameter λ is allowed to change over time according to a Bayesian change-point model with unknown number of change-points. The inclusion of a time-varying autoregressive parameter λ_t is especially suited if outbreak detection is the primary focus, which was not the case in this paper. However, such modifications may also be useful in a retrospective analysis if the infectiousness of a disease changes through public health measures such as increasing vaccination coverage or the reduction of infected person's contact rates by prescribed quarantine. If information on such measures is available, a regression approach might be considered, linking λ_t with the known explanatory variables x_t, e.g. through $\lambda_t = \exp(x_t\beta)$. This would have the advantage

Copyright © 2008 John Wiley & Sons, Ltd.
Statist. Med. 2008; **27**:6250–6267
DOI: 10.1002/sim

that statistical inference is still possible using ML without the need for a fully Bayesian analysis using MCMC.

APPENDIX A: IMPLEMENTATION DETAILS

The log-likelihood given in Section 2.3 needs to be maximized numerically. We use the quasi-Newton method BFGS to obtain the ML estimates and the corresponding standard errors. To ensure positivity of the dispersion parameters ψ_i and the autoregressive parameters λ_i and ϕ_i, these parameters are optimized on the log-scale, i.e. $\tilde{\psi}_i = \log(\psi_i)$, $\tilde{\lambda}_i = \log(\lambda_i)$ and $\tilde{\phi}_i = \log(\phi_i)$ are used instead. The ML estimates and the corresponding standard errors for the original parameters are obtained using the invariance of the ML estimate and the delta method.

In the following, we will give details for the model without unit-specific parameters, i.e. $\tilde{\boldsymbol{\theta}}_i = (\tilde{\lambda}, \tilde{\phi}, \alpha_i, \gamma_1, \ldots, \gamma_S, \delta_1, \ldots, \delta_S)^{\mathrm{T}}$, $i = 1, \ldots, m$. The treatment of the general case is similar. It is not necessary to supply analytic derivatives of the log-likelihood function in `optim` to obtain the ML estimates. However, as the score function

$$s(\tilde{\boldsymbol{\theta}}) = \frac{\partial}{\partial \tilde{\boldsymbol{\theta}}} l(\tilde{\boldsymbol{\theta}}) = \sum_{i,t} \frac{\partial}{\partial \tilde{\boldsymbol{\theta}}} l_{i,t}(\tilde{\boldsymbol{\theta}}_i, \psi)$$

can be derived analytically for our model, we use this information since it speeds up the computation of the ML estimates considerably.

Denote an element of the parameter vector $\tilde{\boldsymbol{\theta}}$ as $\tilde{\theta}_k$; the score function contribution of observation $y_{i,t}$ is then given as

$$\frac{\partial}{\partial \tilde{\theta}_k} l_{i,t}(\tilde{\boldsymbol{\theta}}_i, \tilde{\psi}) = -\frac{\exp(-\tilde{\psi})}{\exp(-\tilde{\psi}) + \mu_{i,t}(\tilde{\boldsymbol{\theta}}_i)} \cdot \frac{\partial}{\partial \tilde{\theta}_k} \mu_{i,t}(\tilde{\boldsymbol{\theta}}_i) + \frac{y}{\mu_{i,t}(\tilde{\boldsymbol{\theta}}_i)} \cdot \frac{\partial}{\partial \tilde{\theta}_k} \mu_{i,t}(\tilde{\boldsymbol{\theta}}_i)$$

$$- \frac{y_{i,t}}{\exp(-\tilde{\psi}) + \mu_{i,t}(\tilde{\boldsymbol{\theta}}_i)} \cdot \frac{\partial}{\partial \tilde{\theta}_k} \mu_{i,t}(\tilde{\boldsymbol{\theta}}_i)$$

$$\frac{\partial}{\partial \tilde{\psi}} l_{i,t}(\tilde{\boldsymbol{\theta}}_i, \tilde{\psi}) = (-\Psi(y_{i,t} + \exp(-\tilde{\psi})) + \Psi(\exp(-\tilde{\psi})) - \log(\exp(-\tilde{\psi})) - 1$$

$$+ \log(\exp(-\tilde{\psi}) + \mu_{i,t}(\tilde{\boldsymbol{\theta}}_i)) + \frac{\exp(-\tilde{\psi}) + y_{i,t}}{\exp(-\tilde{\psi}) + \mu_{i,t}(\tilde{\boldsymbol{\theta}}_i)}) \cdot \exp(-\tilde{\psi})$$

where $\Psi(z) = \mathrm{d}\log(\Gamma(z))/\mathrm{d}z$ is the digamma function [26, p. 258] and

$$\frac{\partial}{\partial \tilde{\lambda}} \mu_{i,t}(\tilde{\boldsymbol{\theta}}_i) = \exp(\tilde{\lambda}) \cdot y_{i,t-1}$$

$$\frac{\partial}{\partial \tilde{\phi}} \mu_{i,t}(\tilde{\boldsymbol{\theta}}_i) = \exp(\tilde{\phi}) \cdot \sum_{j \neq i} w_{ji} y_{j,t-l}$$

$$\frac{\partial}{\partial \alpha} \mu_{i,t}(\tilde{\boldsymbol{\theta}}_i) = v_{i,t}$$

Copyright © 2008 John Wiley & Sons, Ltd.

$$\frac{\partial}{\partial \gamma_s}\mu_{i,t}(\tilde{\boldsymbol{\theta}}_i) = v_{i,t}\cdot\sin(\omega_s t)$$

$$\frac{\partial}{\partial \delta_s}\mu_{i,t}(\tilde{\boldsymbol{\theta}}_i) = v_{i,t}\cdot\cos(\omega_s t)$$

The observed Fisher information matrix $F(\tilde{\boldsymbol{\theta}}) = -\partial s(\tilde{\boldsymbol{\theta}})/\partial\tilde{\boldsymbol{\theta}}$ can also be derived analytically and used to obtain standard errors of the ML estimates.

Newton–Raphson-type methods for optimization require that the initial values are sufficiently close to the solution to guarantee convergence. Far away from the solution, the algorithm can diverge. In addition, it is possible that the optimization algorithm converges to points which are not global optima. Poor convergence or convergence to such local optima occurs much more frequently in multivariate problems involving many parameters than in univariate problems [43]. It is therefore recommended to try multiple starting values to find the global maximum. Even so, results are nearly instantly available.

ACKNOWLEDGEMENTS

We thank the referees for helpful comments and suggestions. Financial support by the Swiss National Science Foundation (SNF) is gratefully acknowledged.

REFERENCES

1. Daley DJ, Gani J. *Epidemic Modelling*: *An Introduction*. Cambridge University Press: Cambridge, 1999.
2. Andersson H, Britton T. *Stochastic Epidemic Models and their Statistical Analysis*. Lecture Notes in Statistics, vol. 151. Springer: New York, 2000.
3. Griffiths DA. Multivariate birth-and-death processes as approximations to epidemic processes. *Journal of Applied Probability* 1973; **10**(1):15–26.
4. Cox DR, Donnelly CA, Bourne FJ, Gettinby G, McInerney JP, Morrison WI, Woodroffe R. Simple model for tuberculosis in cattle and badgers. *Proceedings of the National Academy of Sciences of the United States of America* 2005; **102**(49):17588–17593.
5. Finkenstädt BF, Bjørnstad ON, Grenfell BT. A stochastic model for extinction and recurrence of epidemics: estimation and inference for measles outbreaks. *Biostatistics* 2002; **3**(4):493–510.
6. Finkenstädt BF, Grenfell BT. Time series modelling of childhood diseases: a dynamical systems approach. *Journal of the Royal Statistical Society*, *Series C*: *Applied Statistics* 2000; **49**(2):187–205.
7. Zeger SL, Qaqish B. Markov regression models for time series: a quasi-likelihood approach. *Biometrics* 1988; **44**(4):1019–1031.
8. Knorr-Held L, Richardson S. A hierarchical model for space–time surveillance data on meningococcal disease incidence. *Journal of the Royal Statistical Society*, *Series C*: *Applied Statistics* 2003; **52**(2):169–183.
9. Held L, Höhle M, Hofmann M. A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling* 2005; **5**:187–199.
10. Hament JM, Kimpen JL, Fleer A, Wolfs TF. Respiratory viral infection predisposing for bacterial disease: a concise review. *FEMS Immunology and Medical Microbiology* 1999; **26**(3–4):189–195.
11. Cartwright KA, Jones DM, Smith AJ, Stuart JM, Kaczmarski EB, Palmer SR. Influenza A and meningococcal disease. *The Lancet* 1991; **338**(8766):554–557.
12. Hubert B, Watier L, Garnerin P, Richardson S. Meningococcal disease and influenza-like syndrome: a new approach to an old question. *The Journal of Infectious Diseases* 1992; **166**(3):542–545.
13. Jensen ES, Lundbye-Christensen S, Samuelsson S, Sørensen HT, Schønheyder HC. A 20-year ecological study of the temporal association between influenza and meningococcal disease. *European Journal of Epidemiology* 2004; **19**(2):181–187.

Copyright © 2008 John Wiley & Sons, Ltd.
Statist. Med. 2008; **27**:6250–6267
DOI: 10.1002/sim

14. Jansen AGSC, Sanders EAM, Van Der Ende A, Van Loon AM, Hoes AW, Hak E. Invasive pneumococcal and meningococcal disease: association with influenza virus and respiratory syncytial virus activity? *Epidemiology and Infection* 2008; DOI: 10.1017/S0950268807000271.
15. Robert Koch Institute. SurvStat. Available from: http://www3.rki.de/SurvStat. Accessed July 2007.
16. Farrington CP, Kanaan MN, Gay NJ. Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. *Journal of the Royal Statistical Society*, *Series C*: *Applied Statistics* 2001; **50**:251–283.
17. De Angelis D, Gilks WR, Day NE. Bayesian projection of the acquired immune deficiency syndrome epidemic. *Journal of the Royal Statistical Society*, *Series C*: *Applied Statistics* 1998; **47**:449–481.
18. Brownstein JS, Wolfe CJ, Mandl KD. Empirical evidence for the effect of airline travel on inter-regional influenza spread in the United States. *PLoS Medicine* 2006; **3**(10):e401.
19. Grais RF, Ellis JH, Glass GE. Assessing the impact of airline travel on the geographic spread of pandemic influenza. *European Journal of Epidemiology* 2003; **18**(11):1065–1072.
20. Hufnagel L, Brockmann D, Geisel T. Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences of the United States of America* 2004; **101**(42):15124–15129.
21. Colizza V, Barrat A, Barthélemy M, Vespignani A. The modeling of global epidemics: stochastic dynamics and predictability. *Bulletin of Mathematical Biology* 2006; **68**(8):1893–1921.
22. U.S. Department of Transportation, Bureau of Transportation Statistics. Database: Air Carrier Statistics (Form 41 Traffic): T-100 Domestic Segment (All Carriers). Available from: http://www.transtats.bts.gov. Accessed June 2007.
23. Serfling R. Methods for current statistical analysis of excess pneumonia–influenza deaths. *Public Health Reports* 1963; **78**:494–506.
24. Le Strat Y, Carrat F. Monitoring epidemiologic surveillance data using hidden Markov models. *Statistics in Medicine* 1999; **18**(24):3463–3478.
25. Held L, Hofmann M, Höhle M, Schmid V. A two-component model for counts of infectious diseases. *Biostatistics* 2006; **7**(3):422–437.
26. Abramowitz M, Stegun IA. *Handbook of Mathematical Functions with Formulas*, *Graphs*, *and Mathematical Tables*. Dover: New York, 1964.
27. Thurston SW, Wand MP, Wiencke JK. Negative binomial additive models. *Biometrics* 2000; **56**(1):139–144.
28. Höhle M. Surveillance: an R package for the surveillance of infectious diseases (2007). *Computational Statistics* 2007; **22**(4):571–582.
29. Brundage JF. Interactions between influenza and bacterial respiratory pathogens: implications for pandemic preparedness. *The Lancet Infectious Diseases* 2006; **6**(5):303–312.
30. Schrauder A, Claus H, Elias J, Vogel U, Haas W, Hellenbrand W. Capture–recapture analysis to estimate the incidence of invasive meningococcal disease in Germany, 2003. *Epidemiology and Infection* 2007; **135**(4):657–664.
31. Svetliza CF, Paula GA. Diagnostics in nonlinear negative binomial models. *Communications in Statistics. Theory and Methods* 2003; **32**(6):1227–1250.
32. Diggle PJ. *Time Series. A Biostatistical Introduction*. Oxford University Press: Oxford, 1990.
33. Diggle PJ, Heagerty P, Liang KY, Zeger S. *The Analysis of Longitudinal Data* (*Oxford Statistical Science*). Oxford University Press: Oxford, 2002.
34. Benjamin MA, Rigby RA, Stasinopoulos DM. Generalized autoregressive moving average models. *Journal of the American Statistical Association* 2003; **98**(461):214–223.
35. Davis RA, Dunsmuir WTM, Streett SB. Observation-driven models for Poisson counts. *Biometrika* 2003; **90**(4): 777–790.
36. Centers for Disease Control and Prevention. MMWR Table III (Mortality). Available from: http://www.cdc.gov/EPO/DPHSI/121hist.htm. Accessed December 2006.
37. US Census Bureau, Population Division. Population, population change and estimated components of population change: April 1, 2000 to July 1, 2006 (NST_EST2006_ALLDATA). Available from: http://www.census.gov/popest/datasets.html. Accessed July 2007.
38. Mode CJ. *Multitype Branching Processes—Theory and Applications*. American Elsevier Publishing Company: New York, 1971.
39. Fanshawe T, Diggle P, Rushton S, Sanderson R, Lurz P, Glinianaia S, Pearce M, Parker L, Charlton M, Pless-Mulloli T. Modelling spatio-temporal variation in exposure to particulate matter: a two-stage approach. *Environmetrics* 2008; **19**(6):549–566.

Copyright © 2008 John Wiley & Sons, Ltd.
Statist. Med. 2008; **27**:6250–6267
DOI: 10.1002/sim

40. Dawid AP. Statistical theory. The prequential approach. *Journal of the Royal Statistical Society*, *Series A*: *General* 1984; **147**(2):278–292.
41. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 2007; **102**(477):359–378.
42. Czado C, Gneiting T, Held L. Predictive model assessment for count data. *Technical Report 518*, Department of Statistics, University of Washington, 2007.
43. Thisted RA. *Elements of Statistical Computing. Numerical Computation*. Chapman & Hall: New York, 1988.

Copyright © 2008 John Wiley & Sons, Ltd.
Statist. Med. 2008; **27**:6250–6267
DOI: 10.1002/sim

PAPER III

Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts

Michaela Paul & Leonhard Held

Paper accepted for publication in *Statistics in Medicine*.

Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts

Michaela Paul*, Leonhard Held

Biostatistics Unit, Institute of Social and Preventive Medicine, University of Zurich

Hirschengraben 84, 8001 Zurich, Switzerland

Infectious disease counts from surveillance systems are typically observed in several administrative geographical areas. In this paper, a non-linear model for the analysis of such multiple time series of counts is discussed. To account for heterogeneous incidence levels or varying transmission of a pathogen across regions, region-specific and possibly spatially correlated random effects are introduced. Inference is based on penalized likelihood methodology for mixed models. Since the use of classical model choice criteria such as AIC or BIC can be problematic in the presence of random effects, models are compared by means of one-step-ahead predictions and proper scoring rules. In a case study, the model is applied to monthly counts of meningococcal disease cases in the 94 departments of France (excluding Corsica) and weekly counts of influenza cases in 140 administrative districts of Southern Germany. The predictive performance improves if existing heterogeneity is accounted for by random effects.

Keywords: multivariate time series of counts, infectious diseases, random effects, proper scoring rules

*E-mail: michaela.paul@ifspm.uzh.ch

1. Introduction

Surveillance data on various notifiable diseases collected by national surveillance systems usually consist of multiple time series of counts of new infections. Data for a single disease are typically reported in several strata defined through administrative geographical areas and / or age groups (Giesecke; 2002, Chapter 13). The statistical analysis of the resulting multivariate time series of daily, weekly or monthly counts is an important task in infectious disease epidemiology.

The data are not available at an individual level but aggregated and often also subject to underreporting and reporting delays, which may give rise to overdispersion and blurred dependencies. Thus, detailed (mechanistic) modelling of the epidemic process is not feasible. On the other hand, purely empirical models such as log-linear Poisson regression are not able to adequately capture occasional outbreaks. Such temporal dependence beyond regular patterns can be addressed by including past counts as additional explanatory variables in the predictor (Zeger and Qaqish; 1988; Benjamin et al.; 2003; Fokianos et al.; 2009). In particular, Held et al. (2005) suggest a Poisson regression model with identity link for the mean, i.e. the disease incidence, which is divided into three additive components: the first two components represent an autoregression on past counts in the same and in other regions, respectively, whereas the third component deals with trends and seasonal variation in a usual log-linear formulation. Overdispersion can be allowed for by replacing the Poisson with a negative binomial distribution. Maximum likelihood (ML) estimates are obtained using general optimization routines because the resulting model no longer fits into a generalized linear model (GLM) framework.

When analyzing spatially stratified time series, the assumption of equal transmission rates or incidence levels across all regions is questionable. For example, disease transmission might be influenced by factors such as age, sex, vaccination status, genetic variation in individuals or environmental factors (Becker and Britton; 1999). Such factors could be incorporated into the model framework of Held et al. (2005) as covariates if they are observable and available. Alternatively, to allow for varying transmission of a pathogen across a few regions, region-specific autoregressive parameters are introduced in Paul et al. (2008). However, this approach is limited to a small to moderate number of strata. For highly multivariate time series the estimation procedure can get unstable and identifiability problems may occur.

A common approach to account for unobserved heterogeneity is by means of random effects. For infectious disease data, Li et al. (2003) allowed for household-dependent heterogeneity in infection rates by assuming that the probability of avoiding infection varies randomly in a chain binomial model. Similarly, Davis et al. (2006) considered heterogeneity in transmission probabilities due to household-specific random effects when estimating vaccine efficacy based on outbreak size household data. Zeger and Karim (1991) and Lin and Zhang (1999) used a generalized linear (additive) mixed model for the analysis of longitudinal data on respiratory infection in Indonesian children.

In this paper we introduce random effects in the model discussed in Held et al. (2005) and extended in Paul et al. (2008) to deal with heterogeneous disease transmission and incidence levels. In particular, we focus on how to obtain parameter estimates and on how to compare different models. The estimation of parameters involves integration of the likelihood with respect to the random effects which cannot be done analytically. There are several possible methods for approximating of the integral. For instance, the integral can be solved numerically using adaptive Gauss-Hermite (AGH) quadrature or Monte Carlo integration techniques. Alternatively, the integrand can be approximated such that the integral over the resulting approximation is a closed-form expression. This can be done using first or higher order Laplace approximations or penalized quasi-likelihood (PQL) approaches. See e.g. Breslow (2004) or Tuerlinckx et al. (2006) and references therein for details about these methods in generalized linear mixed models (GLMM).

However, since the model considered in this paper is not linear in parameters, methods for GLMMs cannot be used. Inference is thus based on methodology suggested by Kneib and Fahrmeir (2007) for the analysis of survival data. Estimates of the regression coefficients are derived using penalized likelihood and estimates of variance components are obtained by optimizing the (approximated) marginal likelihood. The estimation procedure corresponds to a variant of the PQL approach as discussed in Breslow and Clayton (1993).

Alternatively, a Bayesian approach could have been used for inference. For example, Held et al. (2006) have proposed a Markov chain Monte Carlo (MCMC) algorithm for a univariate time series model with an autoregressive parameter following a piecewise constant change-point model with unknown number of change-points. This model is extended to a multivariate setting in Hofmann (2007). An extension to a multivariate additive time series model with random effects seems feasible. Knorr-Held and

Richardson (2003) have considered Bayesian inference via MCMC in a somewhat related multivariate model with multiplicative structure. However, model choice in such highly parameterized Bayesian hierarchical models remains a challenge. Of course, the deviance information criterion (DIC) (Spiegelhalter et al.; 2002) can be computed, but recent work (Plummer; 2008) has shown that it tends to prefer too complex models in situations with many random effects.

Similar challenges arise in a frequentist setting where model choice is often based on Akaike's information criterion (Burnham and Anderson; 2002)

$$\mathrm{AIC} = -2\log(L(\hat{\theta}|\mathrm{data})) + 2p\,, \tag{1}$$

where $\log(L(\hat{\theta}|\mathrm{data}))$ is the maximized log-likelihood of a model and p is the number of estimable parameters in this model, or related criteria. Extending the AIC to mixed models raises some issues: First of all, one has to decide on parameters of primary interest in order to choose the appropriate likelihood and the correct number of parameters in equation (1) (Vaida and Blanchard; 2005). In a marginal point of view the marginal likelihood integrated over the random effects is used and the number of estimable parameters is given as the sum of the number of fixed effects and the number of variance parameters. This formulation is appropriate if fixed population effects are of interest. However, if future observations share the same random effects as the observed data (as is the case in our situation), the AIC should be based on the conditional likelihood given the random effects (Vaida and Blanchard; 2005; Greven and Kneib; 2010). The effective number of parameters then lies between that of a model without random effects and that of a model where the random effects are replaced by corresponding fixed effects. Obtaining an unbiased estimator for the effective number of parameters is computationally quite complex already for linear mixed models (Vaida and Blanchard; 2005; Liang et al.; 2008) and not clear for GLMMs. Besides, the use of either the conditional or the marginal AIC for model selection has serious shortcomings e.g. if one wants to decide whether the inclusion of a random effect is necessary or not, i.e. whether the variance is non-zero or zero (Greven and Kneib; 2010).

When using the Bayesian information criterion $\mathrm{BIC} = -2\log(L(\hat{\theta}|\mathrm{data})) + p\log(n)$ the above considerations concerning the choice of an appropriate likelihood and the corresponding number of parameters remain. Furthermore, the effective sample size n is also ambiguous because of the correlation between observations, see e.g. (Pauler; 1998;

Jiang et al.; 2008).

A more natural approach for model selection in time series models than the use of classical model choice criteria is the comparison of successive one-step-ahead forecasts with the actually observed data (Dawid; 1984; Smith; 1985). The usage of proper scoring rules has been recently advocated in this context (Gneiting and Raftery; 2007). The most widely known scoring rule is the logarithmic score which corresponds to the log predictive density at the observed value. In the case of time series where data are inherently ordered, BIC is an approximation of the log marginal likelihood (integrated over unknown parameters), which can be written as the sum of the one-step-ahead logarithmic scores. See Dawid (1984) or Gneiting and Raftery (2007) for further details.

Calculation of successive one-step-ahead forecasts with MCMC is very difficult. Specific algorithms such as particle filters have been suggested (Doucet et al.; 2001) but application to highly multivariate time series data seems not feasible. In contrast, the likelihood-based approach proposed in this paper allows for successive re-fitting the model and calculating the one-step-ahead forecasts.

The rest of the paper is organized as follows. In Section 2 the model described in Paul et al. (2008) is extended for the analysis of highly multivariate time series with heterogeneity in some of the model coefficients by means of (correlated) random effects. Statistical inference based on mixed model methodology is described. Software for fitting the models is currently incorporated in the development version of the `R` package `surveillance` (Höhle; 2007) at `http://surveillance.r-forge.r-project.org/`. The predictive properties are investigated by means of one-step-ahead predictions and proper scoring rules, as outlined in Section 3. In Section 4.1 we apply our model to monthly meningococcal disease counts in the 94 departments of France excluding Corsica from 1985 to 1999, previously analyzed in Knorr-Held and Richardson (2003). The incidence for meningococcal disease is rather low, resulting in a sparse dataset. In a second example in Section 4.2 we consider a more common disease, namely influenza cases observed in the 140 districts of the two German states Baden-Württemberg and Bavaria from 2001–2008. We close with some discussion in Section 5.

2. Model formulation and inference

2.1. Model

To begin, let y_{rt} denote the number of cases of a specific disease in 'unit' $r = 1, \ldots, R$ at time $t = 1, \ldots, T$. Units might represent several geographical regions or different age groups. In the following, we consider spatially stratified counts where a unit corresponds to a certain region. The counts are assumed to be Poisson distributed, $y_{rt}|y_{r,t-1} \sim \text{Po}(\mu_{rt})$, with conditional mean

$$\mu_{rt} = \lambda_r y_{r,t-1} + \phi_r \sum_{q \neq r} w_{qr} y_{q,t-1} + \nu_{rt}\,, \quad \lambda_r, \phi_r, \nu_{rt} > 0\,, \tag{2}$$

see (Paul et al.; 2008). The first two components of μ_{rt} include as covariates the number of past cases at time $t-1$ observed in the same and in other regions, respectively. The weights w_{qr} are assumed to be known and define how cases in other regions relate to cases in region r. In this paper we use $w_{qr} = \mathbb{1}(q \sim r)/n_q$, where $\mathbb{1}$ is the indicator function, $q \sim r$ denotes that region q is a neighbor of region r, and n_q denotes the number of neighbors of region q. See Paul et al. (2008) for a discussion of other possible weights. This observation-driven part of the conditional mean rate should capture occasional outbreaks and, following Held et al. (2005), is called the "epidemic component". The third component ν_{rt} is the "endemic component" and parametrically models seasonal variation and trends.

The specification of the unknown quantities, λ_r, ϕ_r, and ν_{rt}, in the predictor (2) is discussed in more detail later on. At first, we motivate the chosen additive decomposition. There has been much literature on how to generalize the well-known Gaussian autoregressive moving average (ARMA) models to observation-driven models for non-Gaussian time series data, e.g. (Zeger and Qaqish; 1988; Benjamin et al.; 2003; Fokianos et al.; 2009). In the case of a univariate Poisson time series, many models use the GLM framework with log link function and regress $\log(\mu_t)$ on past values of the response to address autocorrelation (e.g. references in Fokianos et al.; 2009). However, it is not possible to account for positive association in a stationary model if past counts are directly included as explanatory variables. Instead, a suitable transformation of y_{t-1} is necessary, (e.g. Benjamin et al.; 2003; Paul et al.; 2008). Models belonging to the class of conditional linear autoregressive models use an identity link and directly regress μ_t on past cases without further transformations. For instance, Fokianos et al. (2009) suggest

the following model for a time series of counts $\{y_t\}$:

$$y_t|\ (\text{History up to time } t) \sim \text{Po}(\mu_t), \quad \mu_t = a\mu_{t-1} + by_{t-1} + d, \quad a, b, d > 0.$$

Without the moving average term $a\mu_{t-1}$, this model represents a univariate version of model (2). It corresponds to a Bienaymé-Galton-Watson branching process with immigration (Haccou et al.; 2007). Branching processes play a fundamental role in the mechanistic modelling of infectious diseases and provide an approximation of the epidemic process assuming an unlimited amount of susceptibles (Farrington et al.; 2003). In a surveillance setting, the number of susceptibles is rarely available and branching processes provide a useful model approximation.

Formulation (2) of our empirical model is based on this branching process formulation and can also be written in a multivariate fashion as

$$\boldsymbol{\mu}_t = \boldsymbol{\Lambda}\boldsymbol{y}_{t-1} + \boldsymbol{\nu}_t \tag{3}$$

with suitably defined column vectors $\boldsymbol{\mu}_t, \boldsymbol{y}_{t-1}$ and $\boldsymbol{\nu}_t$. The matrix $\boldsymbol{\Lambda}$ has entries λ_r on the diagonal and off-diagonal entries $\phi_r w_{qr}$ (Paul et al.; 2008). Stationarity holds if the largest eigenvalue of the matrix $\boldsymbol{\Lambda}$ is smaller than 1 (Held et al.; 2005).

The unknown quantities in (2) are now decomposed additively on the log scale

$$\log(\lambda_r) = \alpha^{(\lambda)} + b_r^{(\lambda)} \tag{4}$$

$$\log(\phi_r) = \alpha^{(\phi)} + b_r^{(\phi)} \tag{5}$$

$$\log(\nu_{rt}) = \alpha^{(\nu)} + b_r^{(\nu)} + \gamma_1 t + \sum_{s=1}^{S} \left\{\gamma_{2s}\sin(\omega_s t) + \gamma_{2s+1}\cos(\omega_s t)\right\} + \log(e_{rt}) \tag{6}$$

where $\alpha^{(\lambda)}, \alpha^{(\phi)}, \alpha^{(\nu)}$ are component-specific intercepts, $b_r^{(\lambda)}, b_r^{(\phi)}, b_r^{(\nu)}$ are random effects, γ_1 is a trend parameter, $\gamma_2, \ldots, \gamma_{2S+1}$ are seasonal parameters, ω_s denotes a specific Fourier frequency, and e_{rt} is a region-specific and time-dependent offset. For instance, $\omega_s = 2\pi s/52$ is used for weekly data (Diggle; 1990). The offset e_{rt} might e.g. comprise yearly varying population numbers. All in all, the vector of the fixed (unpenalized) effects is given by $\boldsymbol{\beta} = (\alpha^{(\lambda)}, \alpha^{(\phi)}, \alpha^{(\nu)}, \gamma_1, \ldots, \gamma_{2S+1})^\top$. However, the parametric modelling of trends and seasonal variation is not restricted to the formulation given in (6). For instance, one could choose a quadratic instead of a linear trend.

Note that to address heterogeneity, the component specific intercepts $\alpha^{(\lambda)}, \alpha^{(\phi)}$, and $\alpha^{(\nu)}$ were allowed to vary across regions in Paul et al. (2008). Such a formulation

is feasible as long as the number of regions is low to moderate. In this paper, each component (4)–(6) instead includes an (optional) random effect. The stacked vector $\boldsymbol{b} = (\boldsymbol{b}^{(\lambda)\top}, \boldsymbol{b}^{(\phi)\top}, \boldsymbol{b}^{(\nu)\top})^\top$ containing all random effects from the three components is assumed to be normally distributed with mean $\mathbf{0}$ and positive definite covariance matrix

$$\boldsymbol{\Sigma} = \text{diag}(\sigma^2_\lambda \mathbf{I}, \sigma^2_\phi \mathbf{I}, \sigma^2_\nu \mathbf{I}) \,, \tag{7}$$

where σ^2_λ, σ^2_ϕ and σ^2_ν are unknown variance parameters, and $\mathbf{I}$ is the $R \times R$ identity matrix. Note that we use the identity matrix for all components, i.e. all elements of $\boldsymbol{b}$ are assumed to be uncorrelated.

However, this assumption might not always be realistic. To allow for correlation between different components, the covariance matrix for the stacked vector $\boldsymbol{b}$ can be specified as

$$\boldsymbol{\Sigma} = \boldsymbol{\Omega} \otimes \mathbf{I} \,, \tag{8}$$

where $\boldsymbol{\Omega}$ is an unknown 3×3 covariance matrix, and $\otimes$ denotes the Kronecker product. Positive definiteness of $\boldsymbol{\Sigma}$ can be ensured by using a suitable parameterization for $\boldsymbol{\Omega}$. See Pinheiro and Bates (1996) for a comparison of several parameterizations for unstructured covariance matrices. We use a computationally stable factorization in terms of standard deviations and correlations which is based on the Cholesky decomposition of $\boldsymbol{\Omega}$ in spherical coordinates, see Appendix A for further details.

In hierarchical models for spatio-temporal data, the independence assumption for a vector of random effects might be questionable and random effects are often assumed to be spatially correlated. As the observation-driven formulation (2) is able to incorporate temporal and spatio-temporal correlation, the inclusion of independent and identically distributed (IID) Gaussian effects in (2) might be sufficient to address possible further heterogeneity. However, the IID random effects within a component might also be replaced by spatially correlated ones.

For example, one might adopt a conditional autoregressive (CAR) model (Besag et al.; 1991; Rue and Held; 2005) for $\boldsymbol{b}^{(\nu)}$, say, i.e. effects in neighboring regions are assumed to be more alike than effects in distant regions. The respective identity matrix in (7) is then replaced by a matrix $\mathbf{K}$ with non-diagonal entries $k_{qr} = -1$ if $q \sim r$, and diagonal entries $k_{qq} = n_q$ (Rue and Held; 2005, p. 102). As the rows and columns of $\mathbf{K}$ sum to zero, this matrix is not of full rank but of rank $g < R$, leading to an improper distribution.

The rank-deficiency of $\mathbf{K}$ equals $R - g = 1$ if all regions are connected and there are no islands. The inverse of $\mathbf{K}$, which is needed to obtain a marginal likelihood estimate for the unknown variance parameter, does thus not exist. However, the vector $\boldsymbol{b}^{(\nu)}$ can be re-expressed via a one-to-one transformation in terms of a $(R-g)$-dimensional fixed, i.e. unpenalized parameter vector and a g-dimensional parameter vector which is IID Gaussian (Kneib and Fahrmeir; 2007; Rue and Held; 2005, p. 91).

Model (2) assumes a Poisson distribution for the counts. This is usually suitable for counts of less prevalent diseases such as e.g. the meningococcal disease data analyzed in Section 4.1. For counts of more prevalent diseases, such as e.g. influenza, overdispersion is more likely (Farrington et al.; 1996). The model formulation can be easily adjusted for overdispersion by replacing the Poisson distribution with the negative binomial, i.e. assuming $y_{rt}|y_{r,t-1} \sim \text{NegBin}(\mu_{rt}, \psi)$, where the conditional mean μ_{rt} is specified exactly as in the Poisson case but the conditional variance increases from μ_{rt} to $\mu_{rt}(1 + \psi\mu_{rt})$ (Held et al.; 2005; Paul et al.; 2008). Thus, the negative binomial model simplifies to the Poisson model for $\psi = 0$. Also note that in applications, each component (4)-(6) of the conditional mean may be omitted in parts or as a whole and the vector $\boldsymbol{\beta}$ then only contains the included fixed effects. Analogously, the vector $\boldsymbol{b}$ only contains the specified random effects with corresponding covariance matrix $\boldsymbol{\Sigma}$.

2.2. Inference

Maximum likelihood estimates in the model formulation without random effects can be obtained by directly maximizing the respective Poisson or negative binomial log-likelihood using numerical optimization routines. See Paul et al. (2008) for further details. In the presence of random effects, inference is more complex. In this paper, we use penalized likelihood approaches (Kneib and Fahrmeir; 2007; Breslow and Clayton; 1993; Schall; 1991; Ogata; 1996; Cai et al.; 2002) to obtain parameter estimates. That is, variance components are treated as fixed when estimating the fixed and random effects. The variance components itself are estimated through maximizing the marginal likelihood after integration with respect to the fixed and random effects. As there is no analytical solution to this integral, a Laplace approximation to the marginal likelihood is used (see e.g. Kneib and Fahrmeir; 2007; Breslow and Clayton; 1993).

In brief, the following alternating algorithm for the estimation of all parameters is used: First update regression coefficients given the current variance parameters, then update

variance components given current regression coefficients via Newton steps. Iterate these two steps until convergence is reached, i.e. parameter estimates do no longer change.

Inference for the regression parameters $\boldsymbol{\beta}, \boldsymbol{b}$, given known variance components, is based on the penalized log-likelihood

$$\ell_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{\Sigma}) = \ell(\boldsymbol{\beta}, \boldsymbol{b}) + \log p(\boldsymbol{b}|\boldsymbol{\Sigma})\,. \tag{9}$$

For simplicity, we assume that the counts are Poisson distributed. The corresponding log-likelihood in (9) is then given by

$$\ell(\boldsymbol{\beta}, \boldsymbol{b}) = \sum_{r,t} y_{rt} \log(\mu_{rt}) - \mu_{rt} - \log \Gamma(y_{rt} + 1)\,,$$

where $\Gamma(\cdot)$ denotes the Gamma function. In the case of overdispersed data, the above Poisson log-likelihood is replaced by the negative binomial log-likelihood, see Paul et al. (2008). The penalty term $\log p(\boldsymbol{b}|\boldsymbol{\Sigma})$ in (9) corresponds to the log distribution of the vector of random effects $\boldsymbol{b}$, i.e. a multivariate Gaussian. After dropping terms that are constant with respect to $\boldsymbol{b}$, we have

$$\log(p(\boldsymbol{b}|\boldsymbol{\Sigma})) = -\frac{1}{2}\boldsymbol{b}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{b}\,.$$

The score vector $\boldsymbol{s}_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{\Sigma})$ can be partitioned into two parts defined by the first-order derivatives of (9) with respect to the fixed and random vectors $\boldsymbol{\beta}$ and $\boldsymbol{b}$. Analogously, the observed Fisher information matrix $\mathbf{F}_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{\Sigma})$, i.e. the negative second-order derivatives of (9), can be partitioned with respect to $\boldsymbol{\beta}$ and $\boldsymbol{b}$. See Appendix A.1 for details. These quantities $\boldsymbol{s}_{\text{pen}}(\cdot)$ and $\mathbf{F}_{\text{pen}}(\cdot)$ allow the computation of updated estimates for the regression coefficients given the variances in a Newton algorithm.

Estimates of the variance components are obtained by maximizing the marginal likelihood

$$L_{\text{marg}}(\boldsymbol{\Sigma}) = \int \exp\Big\{\ell_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{\Sigma})\Big\} d\boldsymbol{\beta} d\boldsymbol{b}. \tag{10}$$

The above integral cannot be solved analytically and one has to resort to numerical methods or approximations. Solving the integral in (10) by means of a Laplace approximation (Tierney and Kadane; 1986) is a relatively simple approach and computationally less complex than other methods mentioned in Section 1. AGH quadrature can provide a more accurate approximation in some situations since the Laplace approximation corresponds to AGH quadrature with a single quadrature point (Liu and Pierce; 1994).

However, the computational complexity of AGH quadrature rapidly increases with the dimension of the random effects vector.

In general, the performance of first order Laplace approximations depends on the validity of the normal approximation. PQL approaches for GLMMs (Breslow and Clayton; 1993) have been found to yield seriously biased estimates of variances and regression parameters with sparse binary outcome data where data are far from normal (Breslow and Lin; 1995; Lin and Breslow; 1996). However, in many situations Laplace approximations work quite well and there is no need for more elaborate approximations. Breslow (2004) investigated several situations and concluded that PQL performs adequately for Poisson GLMMs with mean greater than 5 or even lower for many problems. Kneib and Fahrmeir (2007) showed that the use of a Laplace approximation in their mixed model based inference performed quite similar compared to MCMC based inference for survival data.

Applying a Laplace approximation to the marginal likelihood (10) results in

$$\ell_{\text{marg}}(\boldsymbol{\Sigma}) \approx \ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{b}}) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\hat{\boldsymbol{b}}^\top \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{b}} - \frac{1}{2}\log|\mathbf{F}_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{\Sigma})|,$$

where $|\cdot|$ denotes the determinant of a matrix. Note that both $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}$ are estimated for given $\boldsymbol{\Sigma}$ and thus depend on the variance parameters. In GLMMs (Breslow and Clayton; 1993), generalized additive models (Kneib; 2006) and Cox-type hazard rate models (Kneib and Fahrmeir; 2007), small changes in the variance parameters hardly affect the regression coefficients. We thus assume also that both $\ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{b}})$ and $\hat{\boldsymbol{b}}$ vary only slowly when changing the variance components. Then, the first term of the approximation can be ignored and the log marginal likelihood reduces to

$$\ell_{\text{marg}}(\boldsymbol{\Sigma}) \approx -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\boldsymbol{b}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{b} - \frac{1}{2}\log|\mathbf{F}_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{\Sigma})| \tag{11}$$

where $\boldsymbol{b}$, $\boldsymbol{\beta}$ denote a fixed value not depending directly on the variances, i.e. a current estimate. This expression can be maximized numerically via Newton steps to obtain estimates of the variance parameters.

First and second derivatives of the approximate log marginal likelihood (11) can be derived based on differentiation rules for matrices (see e.g. Harville; 1997). See Appendix A for further details about the estimation procedure and the implementation in the `R` package `surveillance` (Höhle; 2007).

3. Predictive model assessment using proper scoring rules

A classical approach to measure the performance of a model is to use e.g. the mean squared error of several point predictions. However, such an approach does not take into account the uncertainty associated with the point predictions. Instead, probabilistic predictions in the form of a predictive probability distribution P should be considered.

In recent years, the use of strictly proper scoring rules has been advocated to evaluate probabilistic predictions (Gneiting and Raftery; 2007; Czado et al.; 2009). Scoring rules measure the predictive quality by assigning a numerical score, $S(P, y)$, based on a stated predictive distribution P and the later observed true value y. They can be seen as negatively oriented penalties that are to be minimized. A scoring rule is said to be proper, when its expected value under the stated probability distribution P becomes minimal if the observed value y is indeed a realization from P (Gneiting and Raftery; 2007). It is strictly proper if this minimum is unique. Strict propriety is thereby essential to ensure that a scoring rule simultaneously addresses sharpness – the concentration of the predictive distribution – and calibration – the statistical consistency between the predicted and the later observed probability distributions. See Gneiting and Raftery (2007) for further details.

Each scoring rule has differing properties and strengths. Unless there is a clearly defined unique underlying decision problem which calls for a specific scoring rule, it may often be appropriate to investigate several scores in applications (Czado et al.; 2009). In the following, we briefly describe a selection of proper scoring rules for count data discussed in Czado et al. (2009). Denote $\mathrm{P}(Y = k)$ the probability mass function, $\mathrm{P}(Y \leq k)$ the cumulative distribution function and μ_P the first (finite) moment of the predictive probability distribution. Furthermore let y denote the count that materializes.

A traditional summary measure of predictive performance is the squared error score

$$\mathrm{SES}(P, y) = (y - \mu_P)^2$$

defined in analogy to the mean squared error. This scoring rule depends on the predictive distribution only through the first moment μ_P and is proper, but not strictly proper. Nevertheless, we consider this score in Section 4 for comparison, due to its widespread use. Perhaps the most popular strictly proper scoring rule is the logarithmic score (Good;

1952)

$$\text{logS}(P, y) = -\log(\text{P}(Y = y)).$$

The logarithmic score is a local score. It gives no credit for assigning high probabilities to values near but not identical to the count y that materializes. At the same time it is highly sensitive to extreme cases as it strongly penalizes low probability events. A strictly proper scoring rule which is less sensitive to extreme events than the logarithmic score is the ranked probability score (Epstein; 1969; Czado et al.; 2009)

$$\text{RPS}(P, y) = \sum_{k=0}^{\infty} \Big(\text{P}(Y \leq k) - \mathbb{1}(y \leq k) \Big)^2.$$

The RPS adds particular weight to situations with unusually high observed or predicted counts. In such situations, differences between competing models are blown up (Czado et al.; 2009).

Typically, mean scores over a set of predictions are used to rank and compare different models. This is mostly done informally by ordering the obtained mean scores. However, the question whether score differences between models are significant can also be addressed more formally via tests (Diebold and Mariano; 1995; Jolliffe; 2007). Here we will look at a series of one-step-ahead predictions for each considered model. Two models are compared with a Monte Carlo permutation test for paired individual scores (Ludbrook and Dudley; 1998). As test statistic the difference between mean scores from model A and mean scores from model B is used. Both scores are averaged over a set of n individual scores. Under the null hypothesis of no difference, the actually observed difference between mean scores should not be notably different from the distribution of the test statistic under permutation. For each permutation, we first randomly assign the membership of the n individual scores to either model A or B with probability 0.5. We then compute the respective difference in mean for model A and B in this permuted set of scores. As the computation of all possible permutations is only feasible for small datasets, a random sample of permutations is used to obtain the null distribution. The Monte Carlo p-value is then given by

$$\frac{1 + \{\text{number of permuted differences larger than observed difference (in absolute value)}\}}{1 + \text{number of permutations}}.$$

For the model presented in Section 2, the Poisson distribution is used as predictive distribution. The required parameters of the distribution are obtained through plugging

in the estimates of the fixed and random effects. Although such an approach ignores parameter uncertainty, it is the most common and feasible method to obtain a predictive distribution. Fixing the parameters at their estimated values may in exceptional cases lead to identical scores for models of different complexity. For instance, a model with an estimated zero random effects variance would have the same predictions and thus the same scores as a corresponding simpler model without these random effects. In that case a heuristic solution is to always prefer the more parsimonious model among models with equal scores. Such a situation would not occur e.g. in a fully Bayesian approach where the model with random effects would yield a larger predictive variance than the model without these effects and thus lead to different scores.

Note that for a series of one-step-ahead predictions, the model needs to be re-fitted at each time-point. However, the optimizing algorithm usually converges after a few steps when using the previous estimates as initial values because estimates hardly change. A simpler but cruder approach would be not to update the parameter estimates (see e.g. Brix and Diggle; 2001).

4. Applications

In the following we analyze and assess the predictive performance of the proposed model using two spatio-temporal datasets. The first dataset consists of monthly cases of meningococcal disease caused by the *Neisseria meningitidis* bacterium and observed in the 94 departments of France. The incidence of meningococcal disease is low and there is evidence of geographical heterogeneity within France (Knorr-Held and Richardson; 2003). The second dataset comprises the weekly number of laboratory confirmed influenza A and B cases in 140 administrative districts in Southern Germany, obtained from the German national surveillance system operated by the Robert Koch Institute (RKI) (Robert Koch Institute; 2009). The incidence of influenza is very high and the disease spreads across regions in seasonal waves.

In both applications, we consider several models that differ depending on whether autoregressive parameters are included or not, and on whether parameters are treated as fixed or as random. At first, we only consider models with fixed and IID random effects. In the case where there are IID random effects in more than one component, these effects are assumed to be correlated, as this correlated formulation is more flexible than

the formulation with uncorrelated effects. Among all models considered, the best model with lowest mean scores is selected. This model is then compared to a corresponding model with CAR instead of IID effects.

4.1. Meningococcal disease in France

Knorr-Held and Richardson (2003) proposed a complex hierarchical model for the analysis of meningococcal disease incidence. Latent parameters are used to capture temporal, seasonal and spatial trends. Potential autoregressive effects of counts in the same or neighboring regions at previous time points are modulated by latent binary indicators. These indicators are assumed to follow a two-stage hidden Markov model. Thus past counts act multiplicatively on the disease incidence instead of additively as in (2). Fully Bayesian inference is done using MCMC techniques.

In a case study, the model was applied to data on monthly counts of meningococcal disease in the 94 departments of France (excluding Corsica) from 1985 to 1997. Different specifications for the functional form of the autoregressive term, including and excluding spatial neighbors, were compared based on the criterion DIC. Additional data for the years 1998 and 1999 were available for the purpose of prediction but have not been included in the analysis. The main focus was placed on model fit and predictive performance was touched only briefly.

Figure 1 shows the monthly number of cases for the whole of France for the years 1985–1999. A clear seasonal pattern with peaks in winter can be seen. Also shown is a map of the mean yearly incidence per 100 000 habitants averaged over the 15 years. The incidence is regionally varying, with the lowest incidence concentrated in central regions of France. Note that this heterogeneity could be partly due to differences in reporting practices since the data correspond to the compulsory cases notified by clinicians (Knorr-Held and Richardson; 2003). The location of four selected departments that also served as illustration in Knorr-Held and Richardson (2003) are indicated in the map.

Note that in a few regions there seem to have been artefacts in the data collection because a period with no cases at all is followed by a period with regular pattern of cases. To determine such regions we assumed a Poisson-gamma model with a single change-point in mean of unknown location and compared it to a Poisson-gamma model with constant mean (Denison et al.; 2002, Section 3.2.1 and Appendix B.5) for each region. If the posterior probability of the change-point model, assuming equal prior

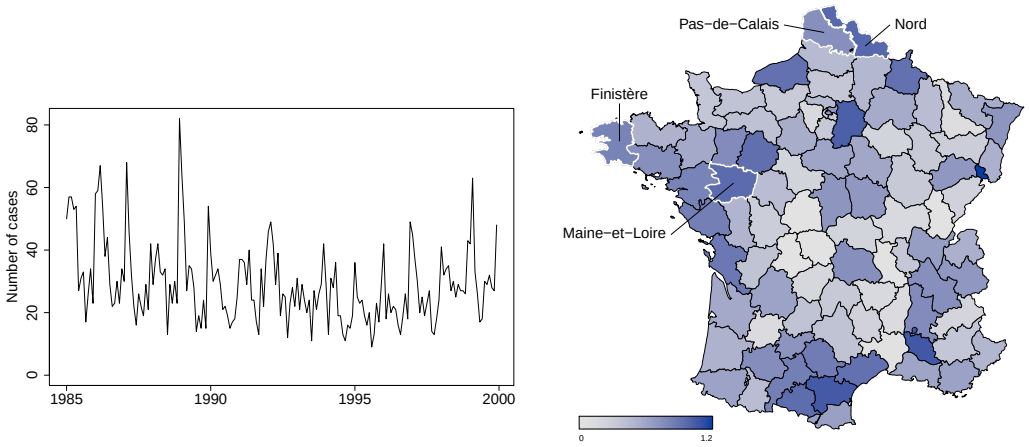


Figure 1: Monthly number of meningococcal disease cases in the whole of France (left) and mean yearly incidence per 100 000 habitants in the 94 departments, averaged over the years 1985 – 1999 (right).

probability for both models, was larger than 0.8, the zero counts for the first month up to the estimated change-point were set to missing for the analysis. Eventually, the first two years in département "Maine-et-Loire" and the first three years in departments "Calvados" and "Marne" were treated as missing.

Since the data are quite sparse we assume a Poisson model. Table 1 summarizes the results for different model formulations. All models contain a quadratic trend, $S = 1$ seasonal components and include expected cases e_{rt} calculated by indirect age-sex standardization as offset in (6). See Knorr-Held and Richardson (2003) for further details on the calculation of expected cases. Table 1 also shows the log-likelihood values for each model. For models without random effects, there is no penalty term involved and accordingly, the marginal likelihood ℓ_{marg} is denoted by NA in the table. In this case, ℓ_{pen} corresponds to the usual Poisson likelihood and could be used for model comparison. In the presence of random effects, both likelihoods are given. However, note that they cannot be used for model comparison.

The first three models in Table 1 contain no autoregression: The intercept $\alpha^{(\nu)}$ is either treated as fixed (denoted by 0 in the label of the model) or as fixed region-specific (1) or as IID (2). The estimates for the quadratic trend and the seasonal components are very similar for all three models and thus not included in the Table. Including the expected cases as offset causes a varying incidence level in model A0. However, it is

Table 1: Analysis of meningococcal disease in the 94 departments of France.

model	$\exp(\hat{\alpha}^{(\lambda)})$ (se)	$\hat{\alpha}^{(\nu)}$ (se)	$\hat{\sigma}^2_\lambda$	$\hat{\sigma}^2_\nu$	$\hat{\rho}_{\lambda\nu}$	maxEV	ℓ_{pen} (ℓ_{mar})
no autoregression							
A0	–	−0.168 (0.022)	–	–	–	–	−10781 (NA)
A1	–	⋆	–	–	–	–	−10431 (NA)
A2	–	−0.253 (0.046)	–	0.14	–	–	−10480 (−146)
fixed intercept $\alpha^{(\lambda)}$							
B1	0.076 (0.009)	⋆	–	–	–	0.08	−10387 (NA)
B2	0.079 (0.009)	−0.327 (0.047)	–	0.14	–	0.08	−10436 (−139)
IID random effects $\boldsymbol{b}^{(\lambda)}$							
C2	0.087 (0.010)	−0.339 (0.047)	0.28	0.14	−0.27	0.22	−10422 (−162)
fixed intercept $\alpha^{(\lambda)}$*, CAR random effects* $\boldsymbol{b}^{(\nu)}$							
D	0.079 (0.009)	−0.328 (0.025)	–	0.50	–	0.08	−10438 (−139)

Shown are parameter estimates with standard errors in brackets. All models contain a quadratic trend and $S = 1$ seasonal term. The maximum eigenvalue of the estimated matrix $\boldsymbol{\Lambda}$ in (3) is denoted by maxEV. Fixed but region-specific intercepts are indicated by ⋆. The numerals in the model labels correspond to 0: fixed intercept $\alpha^{(\nu)}$, 1: region-specific intercept $\alpha_r^{(\nu)}$ 2: IID random effects $\boldsymbol{b}^{(\nu)}$.

not clear whether the heterogeneity of the disease incidence as visible in the map in Figure 1 can thus be captured adequately. Therefore, each region gets its own (fixed) intercept in model A1 which substantially improves the log-likelihood from −10 781 to −10 431. Finally, model A2 assumes that the incidence levels are IID random effects with estimated variance $\hat{\sigma}^2_\nu = 0.14$.

The predictive quality of the models is assessed through one-step-ahead predictions of the last 7 years. Table 2 shows the mean scores based on these 94×84 predictions. The results clearly indicate that a single fixed intercept $\alpha^{(\lambda)}$ (A0) yields the worst predictive performance and regional heterogeneity should be taken into account. According to all scores, a formulation with random effects (A2, B2) is to be preferred compared to the corresponding formulation with region-specific fixed effects (A1, B1). Note that in a regression context it is well-known that the use of plug-in ML or Least Squares estimates

Table 2: Mean scores based on 94×84 one-step-ahead predictions for the meningococcal disease data.

model	$\overline{\text{logS}}$	(p-value)	$\overline{\text{RPS}}$	(p-value)	$\overline{\text{SES}}$	(p-value)
no autoregression						
A0	0.5920	(0.0001)	0.2009	(0.0001)	0.3245	(0.0001)
A1	0.5845	(0.0011)	0.1973	(0.0016)	0.3146	(0.0033)
A2	0.5823	(0.0420)	0.1970	(0.0185)	0.3140	(0.0240)
fixed intercept $\alpha^{(\lambda)}$						
B1	0.5825	(0.0022)	0.1965	(0.0016)	0.3128	(0.0003)
B2	*0.5802*	(0.2260)	*0.1962*	(0.2730)	*0.3121*	(0.0451)
IID random effects $\boldsymbol{b}^{(\lambda)}$						
C2	0.5813	(0.0090)	0.1966	(0.0255)	0.3133	(0.0301)
fixed intercept $\alpha^{(\lambda)}$*, CAR random effects* $\boldsymbol{b}^{(\nu)}$						
D	**0.5799**		**0.1961**		**0.3119**	

Model D is compared to the remaining models, the Monte Carlo p-values are based on permutation tests for paired observations (9999 permutations).

can be suboptimal in prediction problems due to overfitting. Copas (1983, 1997) suggests shrinkage estimates for normal and logistic regression to improve predictive performance.

Basically all models including an autoregression perform better than models excluding autoregression with respect to all scores. The autoregressive parameter $\hat{\alpha}^{(\lambda)}$ in those models is estimated to be quite small, $\exp(\hat{\alpha}^{(\lambda)}) \approx 0.08$. Accordingly, the maximum eigenvalue of $\hat{\boldsymbol{\Lambda}}$ is also small. We did not include the quantity ϕ_r, which measures the influence of adjacent regions, into the model, as it seems to have a very small effect and is hardly estimable from the meningococcal disease dataset.

Note that a model formulation with region-specific fixed autoregressive parameters $\alpha_r^{(\lambda)}$ cannot be fitted while a formulation with random effects $\boldsymbol{b}^{(\lambda)}$ is possible. Figure 2 displays the estimated IID random effects from model C2. The scatterplot shows that there is a moderate negative correlation between $\hat{\boldsymbol{b}}^{(\nu)}$ and $\hat{\boldsymbol{b}}^{(\lambda)}$, in accordance with the estimate $\hat{\rho}_{\lambda\nu} = -0.27$, see Table 1. The estimated autoregressive coefficients $\hat{\lambda}_r$ are also mapped. There is considerable heterogeneity across regions with highest values in some northern and western departments. However, no clear spatial pattern can be seen. Although the variance σ_λ^2 is estimated fairly large in model C2 with IID autoregressive parameter, mean scores for this model are larger than for the respective model B2 with a single fixed parameter.

The best model with lowest mean score among models with only fixed or IID random

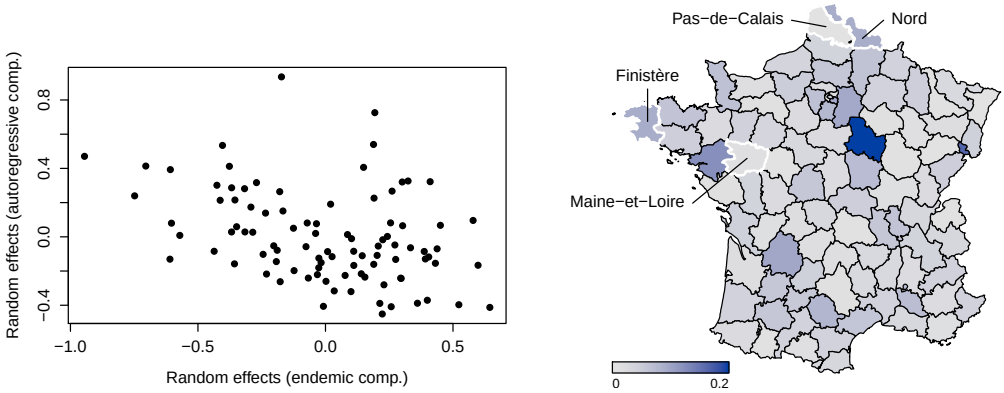


Figure 2: Estimated random effects from model C2. Shown is a scatterplot of the estimated random effects in the endemic component $\hat{\boldsymbol{b}}^{(\nu)}$ and in the autoregressive component $\hat{\boldsymbol{b}}^{(\lambda)}$ (left), and a map of the estimated autoregressive parameters $\hat{\lambda}_r$ (right).

effects is model B2, which includes a single autoregressive parameter $\alpha^{(\lambda)}$ and IID effects $\boldsymbol{b}^{(\nu)}$ in the endemic component. For comparison, we also fitted a corresponding model (D) with CAR instead of IID effects. This formulation leads to a further (if only minor) reduction of all mean scores. Figure 3 shows the estimated mean for this model in the four selected departments. The mean in each département is separated into two additive components: the grey area shows the estimated endemic component $\hat{\nu}_{rt}$ and the blue area corresponds to the autoregressive contribution $\exp(\hat{\alpha}^{(\lambda)})y_{r,t-1}$. The incidence is clearly dominated by the endemic component, which is in agreement with the maximum eigenvalue shown in Table 1.

To judge the differences between the best model D and all competing models more formally, we carried out permutation tests based on 9999 permutations for paired observations. The corresponding Monte Carlo p-values for these tests can be found in Table 2. There is a significant difference to models without autoregression. However, there seems to be no significant difference between the CAR and an IID formulation for the random effects in the endemic component according to both logS and RPS.

A comparison with the results obtained by Knorr-Held and Richardson (2003) is not straightforward due to substantial differences in the model formulation. Also, Knorr-Held and Richardson (2003) have included only data up to 1997 for fitting and have not considered one-step-ahead forecasts for model comparison. Instead, they have used

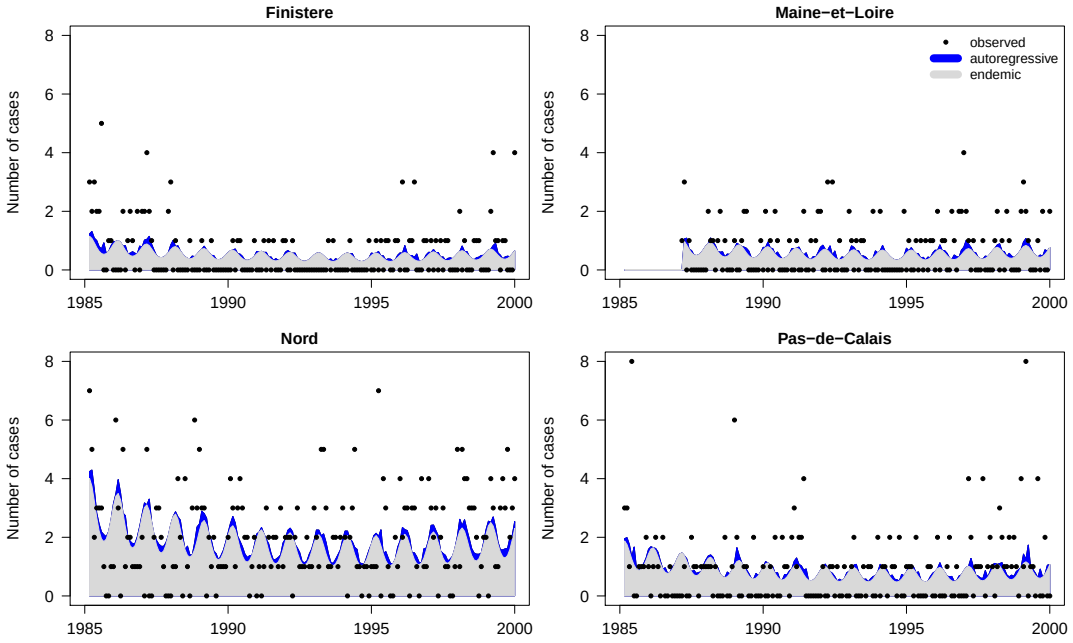


Figure 3: Fitted values for model D for meningococcal disease data in selected departments of France.

DIC (Spiegelhalter et al.; 2002) which has selected a model with two autoregressive terms on the number of counts in the same region and the number of counts in neighboring regions, respectively, as the best model. Note that past counts only enter in a dichotomized form (1 if there was at least one case and 0 otherwise) modulated by additional region-specific latent hidden Markov models. Our approach has identified a significant autoregression on past counts in the same region, but not in neighboring regions.

4.2. Influenza in Southern Germany

As a second application we consider the weekly number of laboratory confirmed influenza cases in 140 administrative districts of the two Southern German states Baden-Württemberg and Bavaria for years 2001 to 2008 (Robert Koch Institute; 2009). Note that the number of regions is too large to fit models with region-specific fixed intercepts. Instead, we apply the proposed methodology with random effects to account for regional heterogeneity.

Figure 4 shows the total number of influenza cases per week in all districts. A clear

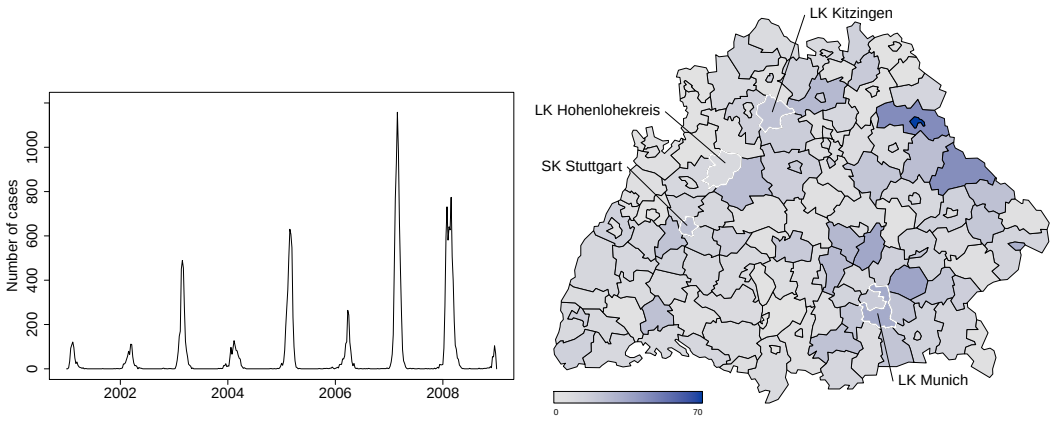


Figure 4: Weekly number of influenza cases in the whole of German states Baden-Württemberg and Bavaria (left) and mean yearly incidence per 100 000 habitants in the 140 administrative districts, averaged over the years 2001 – 2008 (right).

seasonal pattern with steep increases of the number of cases in winter can be seen. Also shown in Figure 4 is a map of the average yearly incidence per 100 000 habitants over the years 2001–2008. The incidence ranges from 0 to 70 cases and is much higher than for the meningococcal disease data. However, note that the influenza data suffer from substantial underreporting. Only a fraction of all influenza cases is actually recorded since the illness caused by influenza is often too slight to warrant medical attention.

We consider several models that differ depending on whether quantities λ_r, ϕ_r are included in the model or not, and if yes, whether they enter as fixed or random effects. The number of cases is assumed to be negative binomial distributed with overdispersion parameter ψ. All models include random incidence levels $\boldsymbol{b}^{(\nu)}$, a linear trend, $S = 3$ seasonal terms and population fractions $p_r / \sum_r p_r$, where p_r denotes the population in region r at 31.12.2001 as offset. Note that $S = 3$ components are used because the superposition of only one sine and cosine wave could not adequately model the distinct peaks, compare Paul et al. (2008).

Table 3 shows estimated parameters and standard errors for the different models. The overdispersion parameter ψ is always estimated to be larger than 1. There seems to be little variation in the autoregressive coefficient λ_r since the variance $\hat{\sigma}^2_\lambda$ is estimated to be quite small. Consequently, estimation of models including IID effects in

Table 3: Analysis of influenza in the 140 administrative districts of Southern Germany.

model	$\exp(\hat{\alpha}^{(\lambda)})$ (se)	$\exp(\hat{\alpha}^{(\phi)})$ (se)	$\hat{\alpha}^{(\nu)}$ (se)	$\hat{\psi}$ (se)	$\hat{\sigma}^2_\lambda$	$\hat{\sigma}^2_\phi$	$\hat{\sigma}^2_\nu$	$\hat{\rho}_{\lambda\nu}$	$\hat{\rho}_{\phi\nu}$	$\hat{\rho}_{\lambda\phi}$	maxEV	ℓ_{pen} (ℓ_{mar})
no autoregression, IID random effects $\boldsymbol{b}^{(\nu)}$												
A0	–	–	0.52 (0.11)	2.07 (0.05)	–	–	0.79	–	–	–	–	−20178 (−314)
A1	–	0.24 (0.01)	0.23 (0.12)	1.80 (0.05)	–	–	1.14	–	–	–	0.24	−19805 (−299)
A2	–	0.39 (0.04)	0.26 (0.11)	1.51 (0.04)	–	1.11	0.79	–	0.51	–	0.39	−19468 (−435)
fixed intercept $\alpha^{(\lambda)}$*, IID random effects* $\boldsymbol{b}^{(\nu)}$												
B0	0.49 (0.02)	–	0.42 (0.10)	1.26 (0.04)	–	–	0.48	–	–	–	0.49	−19049 (−256)
B1	0.46 (0.02)	0.11 (0.01)	0.24 (0.11)	1.20 (0.04)	–	–	0.68	–	–	–	0.57	−18895 (−252)
B2	0.41 (0.02)	0.22 (0.02)	0.22 (0.10)	1.10 (0.03)	–	0.96	0.51	–	0.56	–	0.63	−18742 (−343)
IID random effects $\boldsymbol{b}^{(\lambda)}$*, IID random effects* $\boldsymbol{b}^{(\nu)}$												
C0	0.47 (0.02)	–	0.42 (0.10)	1.22 (0.04)	0.06	–	0.48	0.03	–	–	0.65	−19019 (−301)
C1	0.42 (0.02)	0.11 (0.01)	0.25 (0.11)	1.15 (0.04)	0.09	–	0.67	0.28	–	–	0.67	−18851 (−310)
C2	0.38 (0.02)	0.21 (0.02)	0.22 (0.10)	1.05 (0.03)	0.13	1.02	0.51	0.04	0.53	−0.18	0.75	−18691 (−410)
fixed intercept $\alpha^{(\lambda)}$*, IID random effects* $\boldsymbol{b}^{(\phi)}$*, CAR random effects* $\boldsymbol{b}^{(\nu)}$												
D	0.41 (0.02)	0.22 (0.03)	0.22 (0.06)	1.10 (0.03)	–	1.12	1.66	–	–	–	0.62	−18744 (−368)

Shown are parameter estimates with standard errors in brackets. The maximum eigenvalue of the estimated matrix $\boldsymbol{\Lambda}$ in (3) is denoted by maxEV. All models contain a linear trend and $S = 3$ seasonal terms. The numerals in the model labels correspond to 0: no neighbor-driven component, 1: fixed intercept $\alpha^{(\phi)}$, 2: IID random effects $\boldsymbol{b}^{(\nu)}$.

Table 4: Mean scores based on 140×104 one-step-ahead predictions for the influenza data.

model	$\overline{\text{logS}}$	(p-value)	$\overline{\text{RPS}}$	(p-value)	$\overline{\text{SES}}$	(p-value)
no autoregression, IID random effects $\boldsymbol{b}^{(\nu)}$						
A0	0.5988	(0.0001)	0.5258	(0.0001)	8.1394	(0.0001)
A1	0.5913	(0.0001)	0.5030	(0.0001)	6.7258	(0.0001)
A2	0.5882	(0.0001)	0.4815	(0.0001)	6.2256	(0.0001)
fixed intercept $\alpha^{(\lambda)}$*, IID random effects* $\boldsymbol{b}^{(\nu)}$						
B0	0.5686	(0.0006)	0.4472	(0.0001)	5.2607	(0.6144)
B1	0.5652	(0.0353)	0.4427	(0.0001)	5.1883	(0.9952)
B2	**0.5633**		0.4363		**5.1878**	
IID random effects $\boldsymbol{b}^{(\nu)}$*, IID random effects* $\boldsymbol{b}^{(\nu)}$						
C0	0.5685	(0.0018)	0.4461	(0.0001)	5.2682	(0.5107)
C1	0.5648	(0.0830)	0.4414	(0.0007)	5.2182	(0.7227)
C2	0.5635	(0.5979)	**0.4346**	(0.0438)	5.2088	(0.7399)
fixed intercept $\alpha^{(\lambda)}$*, IID random effects* $\boldsymbol{b}^{(\phi)}$*, CAR random effects* $\boldsymbol{b}^{(\nu)}$						
D	0.5638	(0.1840)	0.4361	(0.5485)	5.1969	(0.4263)

Model B2 is compared to the remaining models, the Monte Carlo p-values are based on permutation tests for paired observations (9999 permutations)

the autoregressive component turned out to be more difficult than estimation of the remaining models in Table 3. On the other hand there is considerable variation concerning the "neighbor-driven" coefficient ϕ_r with $\hat{\sigma}^2_\phi \approx 1$. While the estimated correlation $\hat{\rho}_{\phi\nu}$ in models A2, B2, C2, with IID effects in both the neighbor-driven and endemic component is considerable, IID effects in the autoregressive and endemic component are estimated to be nearly uncorrelated in models C0 and C2. Note that IID and CAR effects in the neighbor-driven and endemic component, respectively, are assumed to be uncorrelated in model D. Both quantities λ_r and ϕ_r characterizing the spatio-temporal spread seem to be important for the model fit. The maximum eigenvalue of $\hat{\mathbf{\Lambda}}$ is quite large in most models which indicates considerable "epidemic" behavior.

The predictive performance of the models is assessed through one-step-ahead predictions of the last 104 weeks, i.e. the last two years. Mean scores based on these 140×104 predictions are shown in Table 4. Most models yield comparable mean scores, except for models without autoregressive and neighbor-driven component (A0, A1, A2) which have a clearly higher score. According to the mean logarithmic score, model B2 with a fixed autoregressive parameter performs best, closely followed by model C2 with IID autoregressive parameter. Both models contain IID effects in the neighbor-driven and in

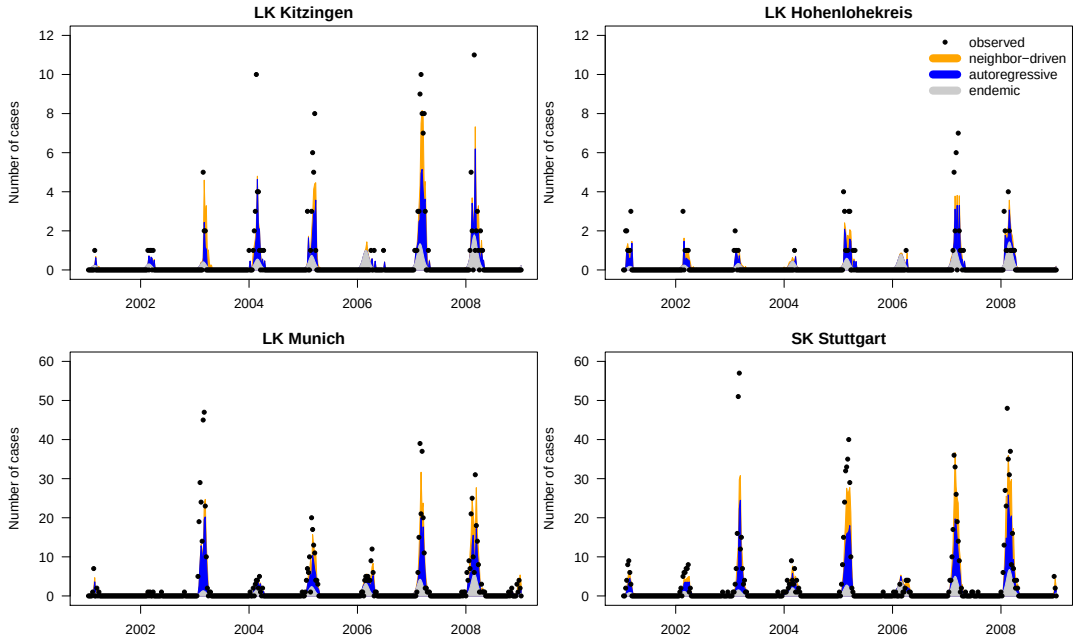


Figure 5: Fitted values for model B2 for influenza data in selected administrative districts of Southern Germany.

the endemic component. Replacing the IID effects in the endemic component of model B2 by CAR effects in model D leads to a slightly higher mean logS. It seems that such a spatial effect is not needed in addition to the neighbor-driven component, which already accounts for spatio-temporal dependence. The ranking of models with respect to $\overline{\text{RPS}}$ is similar except for the best two models with interchanged ordering.

We compared each model to model B2. Models including both autoregressive and neighbor-driven component have a lower mean score (p-values of the Monte Carlo permutation tests < 0.0018 for logS and equal to 0.0001 for RPS). According to the Monte Carlo p-values from the respective permutation tests there is, however, no clear preference for the form of the parameters (i.e. fixed or random).

Figure 5 shows a plot of the fitted incidence from the best model according to $\overline{\text{logS}}$, B2, in four selected administrative districts indicated in the map in Figure 4. Grey, blue and orange areas correspond to the endemic, autoregressive and neighbor-driven part of the mean. In some regions, such as LK Kitzingen or SK Stuttgart, the cases in surrounding districts considerable contribute to the mean incidence. The importance of ϕ_r and λ_r for the predictive performance is also clear when looking at successive maps

of the weekly number of cases.

5. Discussion

In this paper, we have extended the model for the analysis of multiple time series of infectious disease counts suggested by Held et al. (2005) and Paul et al. (2008) to account for different incidence levels or varying disease transmission via possibly correlated random effects. Especially for highly multivariate time series, heterogeneity is very likely to exist and should be accounted for. This is done by including region-specific random effects which are assumed to be either independent, identically distributed Gaussians or follow a conditional autoregressive model. Random effects in the different components of the mean may also be assumed to be correlated.

Inference is based on methodology for mixed models as discussed e.g. in Kneib and Fahrmeir (2007). Estimates for regression parameters are obtained by numerically optimizing the penalized likelihood and estimates of variance parameters are obtained by optimizing an approximate marginal likelihood in an alternating algorithm.

Extension of classical model choice criteria to mixed models is challenging. There are versions of AIC or BIC which can be used to address certain questions, e.g. which fixed effects to select in models with given random effects structure or whether or not to include a random effect in models with given fixed effects structure. However these criteria are not suitable for the comparison of several models in any situation. Therefore, we compared models based on one-step-ahead predictions and proper scoring rules which can be used in any situation.

While a formulation where each region gets its own fixed intercept can be applied for the analysis of disease counts in a low to moderate number of regions, this is no longer feasible when the number of regions is large. For instance, the analysis of the second dataset would not be possible with the formulation in Paul et al. (2008). The analysis of the meningococcal disease and influenza data showed that the induced shrinkage of region-specific estimates towards the overall mean improves the predictive performance in the presence of heterogeneity.

A. Inference

In the following we provide further details about the estimation procedure described in Section 2.2 for the model with mean (2) specified as in Equations (4)-(6), and covariance matrix $\mathbf{\Sigma}$ for the IID random effects given by (8). To ensure positive definiteness of the covariance matrix, we use a parameterization based on the Cholesky decomposition of $\mathbf{\Omega}$ in spherical coordinates (Pinheiro and Bates; 1996). Note that the diagonal elements of the lower triangular Cholesky matrix must be positive and the off-diagonal elements must be in $(0, \pi)$ to ensure uniqueness. These restrictions can be guaranteed by using suitable transformations (see e.g. Rapisarda et al.; 2007). The resulting parameterization is then given by

$$\mathbf{\Omega}(\boldsymbol{\theta}) = \begin{pmatrix} \exp(2s_1) & & \\ \frac{r_1 \exp(s_2+s_1)}{\sqrt{r_1^2+1}} & \exp(2s_2) & \\ \frac{r_2 \exp(s_3+s_1)}{\sqrt{r_2^2+1}} & \frac{\left(r_1 r_2 \sqrt{r_3^2+1}+r_3\right) \exp(s_3+s_2)}{\sqrt{r_1^2+1}\sqrt{r_2^2+1}\sqrt{r_3^2+1}} & \exp(2s_3) \end{pmatrix}$$

where the vector $\boldsymbol{\theta} = (s_1, s_2, s_3, r_1, r_2, r_3)^\top$ contains the unknown variance parameters which can take values on the whole real line. For the sake of clarity only the lower triangular of the symmetric matrix $\mathbf{\Omega}(\boldsymbol{\theta})$ is displayed. Note that the diagonal elements correspond to the variances of the random effects in the autoregressive (4), neighbor-driven (5), and endemic component (6). For instance, we have $\sigma_\lambda^2 = \exp(2s_1)$. Similarly, we obtain the correlation parameters, e.g. $\rho_{\lambda\phi} = r_1/\sqrt{r_1^2+1}$.

A.1. Estimation of regression parameters

First, let i and j denote any element of the stacked vector of fixed and random effects. The score vector with first derivatives of (9) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{b}$ is given by

$$\boldsymbol{s}_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b}; \mathbf{\Sigma}) = \begin{pmatrix} \frac{\partial \ell_{\text{pen}}(\boldsymbol{\beta},\boldsymbol{b})}{\partial \boldsymbol{\beta}} \\ \\ \frac{\partial \ell_{\text{pen}}(\boldsymbol{\beta},\boldsymbol{b})}{\partial \boldsymbol{b}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{s}(\boldsymbol{\beta}) \\ \\ \boldsymbol{s}(\boldsymbol{b}) - \mathbf{\Sigma}^{-1}\boldsymbol{b} \end{pmatrix},$$

where the elements of the unpenalized score vector are given by

$$\boldsymbol{s}(i) = \sum_{r,t} \frac{y_{rt}}{\mu_{rt}} \frac{\partial \mu_{rt}}{\partial i} - \frac{\partial \mu_{rt}}{\partial i}.$$

Analogously, the observed Fisher information matrix is partitioned as

$$\mathbf{F}_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{\Sigma}) = \begin{pmatrix} -\frac{\partial^2 \ell_{\text{pen}}(\boldsymbol{\beta},\boldsymbol{b})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} & -\frac{\partial^2 \ell_{\text{pen}}(\boldsymbol{\beta},\boldsymbol{b})}{\partial \boldsymbol{\beta} \partial \boldsymbol{b}^\top} \\ -\frac{\partial^2 \ell_{\text{pen}}(\boldsymbol{\beta},\boldsymbol{b})}{\partial \boldsymbol{b} \partial \boldsymbol{\beta}^\top} & -\frac{\partial^2 \ell_{\text{pen}}(\boldsymbol{\beta},\boldsymbol{b})}{\partial \boldsymbol{b} \partial \boldsymbol{b}^\top} \end{pmatrix} = \begin{pmatrix} \mathbf{F}[\boldsymbol{\beta}\boldsymbol{\beta}] & \mathbf{F}[\boldsymbol{\beta}\boldsymbol{b}] \\ \mathbf{F}[\boldsymbol{\beta}\boldsymbol{b}] & \mathbf{F}[\boldsymbol{b}\boldsymbol{b}] + \boldsymbol{\Sigma}^{-1} \end{pmatrix}, \tag{12}$$

where $\mathbf{F}[\boldsymbol{i}\boldsymbol{j}]$ denotes the block of the unpenalized Fisher information matrix corresponding to the parameter vectors $\boldsymbol{i}$ and $\boldsymbol{j}$, and has elements

$$\mathbf{F}[ij] = -\sum_{r,t} \left\{ -\frac{y_{rt}}{\mu_{rt}^2} \frac{\partial \mu_{rt}}{\partial i} \frac{\partial \mu_{rt}}{\partial j} + \frac{y_{rt}}{\mu_{rt}} \frac{\partial^2 \mu_{rt}}{\partial i \partial j} - \frac{\partial^2 \mu_{rt}}{\partial i \partial j} \right\}.$$

In the following, let $\boldsymbol{\xi}^{(\lambda)}$ contain all parameters in the autoregressive component (4), i.e. $\alpha^{(\lambda)}, \boldsymbol{b}^{(\lambda)}$, and let $\boldsymbol{x}_{rt}^{(\lambda)}$ denote the corresponding design vector in region r at time t, i.e. $\boldsymbol{x}_{rt}^{(\lambda)} = (1, \boldsymbol{u}_r)^\top$ where the r-th element of the R-dimensional vector $\boldsymbol{u}_r$ is 1 and 0 otherwise. The vectors $\boldsymbol{\xi}^{(\phi)}$, $\boldsymbol{x}_{rt}^{(\phi)}$ and $\boldsymbol{\xi}^{(\nu)}$, $\boldsymbol{x}_{rt}^{(\nu)}$ are defined in analogy for components (5) and (6), respectively. First and second partial derivatives of μ_{rt} are then given by

$$\frac{\partial \mu_{rt}}{\partial \boldsymbol{\xi}^{(\lambda)}} = \lambda_{rt} y_{r,t-1} \, \boldsymbol{x}_{rt}^{(\lambda)}, \quad \frac{\partial \mu_{rt}}{\partial \boldsymbol{\xi}^{(\phi)}} = \phi_{rt} \sum_{q \neq r} w_{qr} y_{q,t-1} \, \boldsymbol{x}_{rt}^{(\phi)}, \quad \frac{\partial \mu_{rt}}{\partial \boldsymbol{\xi}^{(\nu)}} = \nu_{rt} \, \boldsymbol{x}_{rt}^{(\nu)}$$

and

$$\frac{\partial^2 \mu_{rt}}{\partial \boldsymbol{\xi}^{(\lambda)} \partial \boldsymbol{\xi}^{(\lambda)\top}} = \lambda_{rt} y_{r,t-1} \, \boldsymbol{x}_{rt}^{(\lambda)} \boldsymbol{x}_{rt}^{(\lambda)\top}, \qquad \frac{\partial^2 \mu_{rt}}{\partial \boldsymbol{\xi}^{(\phi)} \partial \boldsymbol{\xi}^{(\phi)\top}} = \phi_{rt} \sum_{q \neq r} w_{qr} y_{q,t-1} \, \boldsymbol{x}_{rt}^{(\phi)} \boldsymbol{x}_{rt}^{(\phi)\top},$$

$$\frac{\partial^2 \mu_{rt}}{\partial \boldsymbol{\xi}^{(\nu)} \partial \boldsymbol{\xi}^{(\nu)\top}} = \nu_{rt} \, \boldsymbol{x}_{rt}^{(\nu)} \boldsymbol{x}_{rt}^{(\nu)\top}.$$

Second partial derivatives of μ_{rt} with respect to parameters belonging to different components are zero.

A.2. Estimation of variance parameters

First note that

$$\log |\boldsymbol{\Sigma}| = \log(|\boldsymbol{\Omega}(\boldsymbol{\theta}) \otimes \mathbf{I}|) = \log(|\boldsymbol{\Omega}(\boldsymbol{\theta})|^R |\mathbf{I}|^3) = 2R \left(\sum_{i=1}^{3} s_i - \frac{1}{2} \sum_{i=1}^{3} \log(r_i^2 + 1) \right).$$

Let k and l denote any element of the vector with variance parameters $\boldsymbol{\theta}$. Furthermore, the dependence of the observed Fisher information matrix (12) on both regression and variance parameters will be suppressed in the following for notational convenience. The score vector $\boldsymbol{s}_{\text{marg}}(\boldsymbol{\Sigma})$ with first derivatives of (11) with respect to $\boldsymbol{\theta}$ then has elements

$$\frac{\partial \ell_{\text{marg}}(\boldsymbol{\Sigma})}{\partial k} = -R \sum_{i=1}^{3} \left(\mathbb{1}(k = s_i) - \frac{r_i}{r_i^2 + 1} \mathbb{1}(k = r_i) \right) + \frac{1}{2} \boldsymbol{b}^\top \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial k} \boldsymbol{\Sigma}^{-1} \boldsymbol{b} - \frac{1}{2} \operatorname{tr} \left(\mathbf{F}_{\text{pen}}^{-1} \frac{\partial \mathbf{F}_{\text{pen}}}{\partial k} \right)$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. The elements of the observed Fisher information matrix $\mathbf{F}_{\text{marg}}(\boldsymbol{\Sigma})$ are given as

$$\begin{aligned}\frac{\partial^2 \ell_{\text{marg}}(\boldsymbol{\Sigma})}{\partial k \partial l} =& - R \sum_{i=1}^{3} \left(\frac{r_i^2 - 1}{(r_i^2 + 1)^2} \mathbb{1}(k = l = r_i) \right) \\ & - \frac{1}{2} \boldsymbol{b}^\top \boldsymbol{\Sigma}^{-1} \left(- \frac{\partial^2 \boldsymbol{\Sigma}}{\partial k \partial l} + \frac{\partial \boldsymbol{\Sigma}}{\partial k} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial l} + \frac{\partial \boldsymbol{\Sigma}}{\partial l} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial k} \right) \boldsymbol{\Sigma}^{-1} \boldsymbol{b} \\ & - \frac{1}{2} \mathrm{tr} \left(- \mathbf{F}_{\text{pen}}^{-1} \frac{\partial \mathbf{F}_{\text{pen}}}{\partial l} \mathbf{F}_{\text{pen}}^{-1} \frac{\partial \mathbf{F}_{\text{pen}}}{\partial k} + \mathbf{F}_{\text{pen}}^{-1} \frac{\partial^2 \mathbf{F}_{\text{pen}}}{\partial k \partial l} \right) .\end{aligned}$$

The derivation of first and second derivatives of $\boldsymbol{\Omega}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, and thus $\boldsymbol{\Sigma}$ and $\mathbf{F}_{\text{pen}}$, is straightforward and omitted here. Note that for uncorrelated random effects, the block diagonal structure of the covariance matrix $\boldsymbol{\Sigma}$ can be exploited to obtain simplified expressions for the score vector $\boldsymbol{s}_{\text{marg}}(\boldsymbol{\Sigma})$ and the Fisher information matrix $\mathbf{F}_{\text{marg}}(\boldsymbol{\Sigma})$.

A.3. Implementation

In applications, the likelihood functions described in Section 2 might be flat and contorted due to the additive decomposition of the mean. This complicates maximization. In principle, estimates of both regression and variance components can be obtained using the iterative Newton algorithm. At each iteration, the score function is approximated by a local quadratic model and solved for the root, i.e. the maximum of the likelihood. Close to the maximum, the Fisher information matrix must be positive definite and the Newton method guarantees fast convergence (Thisted; 1988). However, far away from the maximum the Fisher information matrix might not be positive definite. If initial values are poor, the Newton algorithm may fail.

There are several modifications to obtain a globally convergent Newton algorithm without sacrificing the good local convergence properties, see e.g. Dennis and Schnabel (1996, Chapters 5, 6). For instance, the Fisher information matrix can be modified such that it is always positive definite. Convergence from poor initial values can be improved by using a trust-region approach (Dennis and Schnabel; 1996, Chapter 6.4). The trust region approach provides at each iteration a region in which the quadratic approximation can be trusted to adequately model the score function. The root of the approximation is picked as new iterate and the goodness of the approximation is evaluated. If the approximation is not good enough, the region is shrunken until an acceptable new iterate is found.

The optimization problem gets harder the more parameters are involved and the choice of good initial values is especially important. In nonlinear models, it is often advisable to use several initial values and select estimates resulting in the highest likelihood should the optimization algorithm lead to different results (e.g. Paul et al.; 2008). In the current context, however, it is not clear how to deal with differing results for the alternating algorithm from Section 2.2. The penalized likelihood can not be used for deciding between two sets of parameter estimates because it depends on the variance parameters. Thus, results with larger variance parameters tend to be preferred.

The estimation procedure is currently incorporated in the development version of the `R` package `surveillance` (Höhle; 2007) available from `http://surveillance.r-forge.r-project.org/`. We use the `R` function `nlminb` as default optimizer which can either use analytical second derivatives or numerical approximations to the Hessian. Although a single computation of the analytical Fisher information matrix is more expensive than the use of numerical approximations, the use of the analytical Fisher information matrix makes convergence faster in the long run and is to be preferred. The function `nlminb` also implements a trust region approach which enhances convergence. In all applications considered, the algorithm always lead to the same solution after convergence when using different initial values.

Acknowledgments

We thank the reviewers for helpful comments and suggestions. Financial support by the Swiss National Science Foundation is gratefully acknowledged.

References

Becker, N. G. and Britton, T. (1999). Statistical studies of infectious disease incidence, *Journal of the Royal Statistical Society. Series B* **61**(2): 287–307.

Benjamin, M. A., Rigby, R. A. and Stasinopoulos, D. M. (2003). Generalized autoregressive moving average models, *Journal of the American Statistical Association* **98**(461): 214–223.

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration with two applica-

tions in spatial statistics, *Annals of the Institute of Statistical Mathematics* **43**(1): 1–20.

Breslow, N. E. (2004). Whither PQL?, *in* Lin, D Y and Heagerty, P J (ed.), *Proceedings of the Second Seattle Symposium in Biostatistics - Analysis of Correlated Data*, Vol. 179 of *Lecture Notes in Statistics*, Springer, New York, pp. 1–22.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**(421): 9–25.

Breslow, N. E. and Lin, X. H. (1995). Bias correction in generalized linear mixed models with a single component of dispersion, *Biometrika* **82**(1): 81–91.

Brix, A. and Diggle, P. J. (2001). Spatiotemporal prediction for log-Gaussian Cox processes, *Journal of the Royal Statistical Society. Series B* **63**(4): 823–841.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach*, 2nd edn, Springer, New York.

Cai, T., Hyndman, R. J. and Wand, M. (2002). Mixed model-based hazard estimation, *Journal of Computational and Graphical Statistics* **11**(4): 784–798.

Copas, J. B. (1983). Regression, prediction and shrinkage, *Journal of the Royal Statistical Society. Series B* **45**(3): 311–354.

Copas, J. B. (1997). Using regression models for prediction: shrinkage and regression to the mean, *Statistical Methods in Medical Research* **6**(2): 167–183.

Czado, C., Gneiting, T. and Held, L. (2009). Predictive model assessment for count data, *Biometrics* **65**(4): 1254–1261.

Davis, X., Waller, L. and Haber, M. (2006). Estimating vaccine efficacy from outbreak size household data in the presence of heterogeneous transmission probabilities, *Journal of Biopharmaceutical Statistics* **16**(4): 499–516.

Dawid, A. P. (1984). Statistical theory. The prequential approach, *Journal of the Royal Statistical Society. Series A.* **147**(2): 278–292.

Denison, D. G. T., Holmes, C. C., Mallick, B. K. and Smith, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*, Wiley Series in Probability and Statistics, John Wiley & Sons Ltd., Chichester.

Dennis, J. E. and Schnabel, R. B. (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Society for Industrial and Applied Mathematics, Philadelphia.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy, *Journal of Business & Economic Statistics* **13**(3): 253–263.

Diggle, P. J. (1990). *Time Series. A Biostatistical Introduction*, Oxford University Press, Oxford.

Doucet, A., de Freitas, N. and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*, Statistics for Engineering and Information Science, 2 edn, Springer, Berlin.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories, *Journal of Applied Meteorology* **8**(6): 985–987.

Farrington, C. P., Andrews, N. J., Beale, A. D. and Catchpole, M. A. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease, *Journal of the Royal Statistical Society. Series A* **159**: 547–563.

Farrington, C. P., Kanaan, M. N. and Gay, N. J. (2003). Branching process models for surveillance of infectious diseases controlled by mass vaccination, *Biostatistics* **4**(2): 279–295.

Fokianos, K., Rahbek, A. and Tjøstheim, D. (2009). Poisson autoregression, *Journal of the American Statistical Association* **104**(488): 1430–1439.

Giesecke, J. (2002). *Modern Infectious Disease Epidemiology*, 2 edn, Hodder Arnold, London.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association* **102**(477): 359–378.

Good, I. J. (1952). Rational decisions, *Journal of the Royal Statistical Society. Series B* **14**(1): 107–114.

Greven, S. and Kneib, T. (2010). On the behavior of marginal and conditional Akaike information criteria in linear mixed models, *Biometrika* . doi:10.1093/biomet/asq042.

Haccou, P., Jagers, P. and Vatutin, V. A. (2007). *Branching Processes: Variation, Growth, and Extinction of Populations*, Cambridge University Press, Cambridge.

Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*, Springer, New York.

Held, L., Hofmann, M., Höhle, M. and Schmid, V. (2006). A two-component model for counts of infectious diseases, *Biostatistics* **7**(3): 422–437.

Held, L., Höhle, M. and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts, *Statistical Modelling* **5**(3): 187–199.

Hofmann, M. (2007). *Statistical Models for Infectious Disease Surveillance Counts*, PhD thesis, Ludwig-Maximilians-Universität Munich. `http://edoc.ub.uni-muenchen.de/6601/`.

Höhle, M. (2007). Surveillance: an R package for the monitoring of infectious diseases, *Computational Statistics* **22**(4): 571–582.

Jiang, J. M., Rao, J. S., Gu, Z. H. and Nguyen, T. (2008). Fence methods for mixed model selection, *Annals of Statistics* **36**(4): 1669–1692.

Jolliffe, I. T. (2007). Uncertainty and inference for verification measures, *Weather and Forecasting* **22**(3): 637–650.

Kneib, T. (2006). *Mixed Model based Inference in Structured Additive Regression*, PhD thesis, LMU München: Fakultät für Mathematik, Informatik und Statistik. `http://edoc.ub.uni-muenchen.de/5011/`.

Kneib, T. and Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression, *Scandinavian Journal of Statistics* **34**(1): 207–228.

Knorr-Held, L. and Richardson, S. (2003). A hierarchical model for space-time surveillance data on meningococcal disease incidence, *Journal of the Royal Statistical Society. Series C* **52**(2): 169–183.

Li, N., Qian, G. Q. and Huggins, R. (2003). A random effects model for diseases with heterogeneous rates of infection, *Journal of Statistical Planning and Inference* **116**(1): 317–332.

Liang, H., Wu, H. L. and Zou, G. H. (2008). A note on conditional AIC for linear mixed-effects models, *Biometrika* **95**(3): 773–778.

Lin, X. H. and Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion, *Journal of the American Statistical Association* **91**(435): 1007–1016.

Lin, X. H. and Zhang, D. W. (1999). Inference in generalized additive mixed models by using smoothing splines, *Journal of the Royal Statistical Society. Series B* **61**(2): 381–400.

Liu, Q. and Pierce, D. A. (1994). A note on Gauss-Hermite quadrature, *Biometrika* **81**(3): 624–629.

Ludbrook, J. and Dudley, H. (1998). Why permutation tests are superior to t and F tests in biomedical research, *American Statistician* **52**(2): 127–132.

Ogata, Y. (1996). Evaluation of spatial Bayesian models - two computational methods, *Journal of Statistical Planning and Inference* **51**(1): 1–18.

Paul, M., Held, L. and Toschke, A. M. (2008). Multivariate modelling of infectious disease surveillance data, *Statistics in Medicine* **27**(29): 6250–6267.

Pauler, D. K. (1998). The Schwarz criterion and related methods for normal linear models, *Biometrika* **85**(1): 13–27.

Pinheiro, J. C. and Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices, *Statistics and Computing* **6**(3): 289–296.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison, *Biostatistics* **9**(3): 523–539.

Rapisarda, F., Brigo, D. and Mercurio, F. (2007). Parameterizing correlations: a geometric interpretation, *IMA Journal of Management Mathematics* **18**(1): 55–73.

Robert Koch Institute (2009). SurvStat, `http://www3.rki.de/SurvStat`. Accessed March 2009.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*, Vol. 104, Chapman & Hall/CRC Press, London.

Schall, R. (1991). Estimation in generalized linear models with random effects, *Biometrika* **78**(4): 719–727.

Smith, J. Q. (1985). Diagnostic checks of non-standard time series models, *Journal of Forecasting* **4**(3): 283–291.

Spiegelhalter, D. J., Best, N. G., Carlin, B. R. and van der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society. Series B* **64**: 583–616.

Thisted, R. A. (1988). *Elements of Statistical Computing. Numerical Computation*, Chapman and Hall, New York.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association* **81**(393): 82–86.

Tuerlinckx, F., Rijmen, F., Verbeke, G. and De Boeck, P. (2006). Statistical inference in generalized linear mixed models: a review, *British Journal of Mathematical and Statistical Psychology* **59**(2): 225–255.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models, *Biometrika* **92**(2): 351–370.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects: a Gibbs sampling approach, *Journal of the American Statistical Association* **86**(413): 79–86.

Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach, *Biometrics* **44**(4): 1019–1031.

PAPER IV

Heterogeneity in vaccination coverage explains the size and occurrence of measles epidemics in German surveillance data

Sereina A. Herzog, Michaela Paul & Leonhard Held

Paper published in *Epidemiology and Infection*, 2010 (in press).

Epidemiol. Infect., Page 1 of 11. © Cambridge University Press 2010
doi:10.1017/S0950268810001664

Heterogeneity in vaccination coverage explains the size and occurrence of measles epidemics in German surveillance data

S. A. HERZOG[1]*†, M. PAUL[2]† AND L. HELD[2]

[1] *Institute of Social and Preventive Medicine, University of Bern, Switzerland*
[2] *Biostatistics Unit, Institute of Social and Preventive Medicine, University of Zurich, Switzerland*

(*Accepted 15 June 2010*)

SUMMARY

The objective of this study was to characterize empirically the association between vaccination coverage and the size and occurrence of measles epidemics in Germany. In order to achieve this we analysed data routinely collected by the Robert Koch Institute, which comprise the weekly number of reported measles cases at all ages as well as estimates of vaccination coverage at the average age of entry into the school system. Coverage levels within each federal state of Germany are incorporated into a multivariate time-series model for infectious disease counts, which captures occasional outbreaks by means of an autoregressive component. The observed incidence pattern of measles for all ages is best described by using the log proportion of unvaccinated school starters in the autoregressive component of the model.

Key words: Infectious disease epidemiology, measles (rubeola), MMR vaccination, modelling.

INTRODUCTION

Measles is a highly contagious disease and still an important health concern [1]. Numerous efforts such as routine childhood vaccination programmes or the WHO measles elimination plan have significantly reduced the incidence of measles in Europe. The epidemic pattern has changed from a roughly biennial cycle to an irregular sequence of outbreaks [2]. However, disease has not been eradicated. The incidence of measles varies widely, with large outbreaks in Romania, Germany, UK, Switzerland and Italy in 2006 and 2007, whereas in other countries such as Finland, Slovakia and Hungary almost no cases were reported [1]. Since most measles cases were unvaccinated or incompletely vaccinated, the differences in incidence are likely to be due to differences in the success of national vaccination programmes [1, 2]. For instance, there have been several outbreaks in some of the 16 federal states of Germany in recent years [3–6]. Detailed investigations of selected outbreaks showed that most cases occurred in unvaccinated individuals [4].

National surveillance systems such as that at the Robert Koch Institute (RKI), Germany, typically provide weekly time-series of counts stratified by for example, region, age or sex. Accordingly, statistical methods for the analysis of multivariate time-series of counts are needed. It is of public health interest to investigate empirically the relationship between vaccination coverage and the occurrence and size of measles epidemics using such data.

* Author for correspondence: Ms. S. Herzog, Institute of Social and Preventive Medicine, University of Bern, Finkenhubelweg 11, 3012 Bern, Switzerland.
(Email: sherzog@ispm.unibe.ch)
† These authors contributed equally to this work.

Cummings *et al.* [7] used a linear model to analyse the sum of measles cases over 5 years in several provinces of Cameroon, including vaccination coverage among other covariates. However, the time-series aspect was not considered. Multivariate time-series methods for counts of infectious diseases have only recently been developed and applied to epidemiological data. However, these models are not able to cope with occasional large outbreaks. For instance, Frank *et al.* [8] investigated the association between human infection with Shiga toxin-producing *Escherichia coli* (STEC) and cattle density based on German notification data. A Bayesian Poisson regression model was used to analyse the weekly number of cases in each age group and district of Germany. The model accounted for temporal and seasonal trends, spatial variation and cattle density as explanatory factors. No large STEC gastroenteritis outbreaks occurred in the time period considered. Hens *et al.* [9] modelled the yearly, age-stratified incidence of hepatitis B in Bulgaria using a log-additive Poisson model, where age and time were modelled as non-parametric functions. The impact of vaccination was taken into account by including indicators for various immunization programmes as covariates. The log-additive Poisson model chosen was justified since the data contained no outbreaks.

If there are outbreaks in the data, a more realistic formulation for (multivariate) time-series of infectious disease counts has been suggested by Held *et al.* [10]. The model decomposes the disease incidence into two additive components. One component represents an autoregression on past counts which allows for temporal dependence beyond regular patterns, i.e. epidemic behaviour. The other component accounts for regular, endemic behaviour. However, this method did not consider the inclusion of covariates.

The aim of this paper is to investigate the association between vaccination coverage and the size and occurrence of measles epidemics. We first describe the data about measles incidence [11] and vaccination coverage [12] in Germany obtained from the RKI. The approach of Held *et al.* [10] is extended to allow for the inclusion of covariates and applied to the measles data using vaccination coverage as an explanatory variable. Different formulations of the proposed model are compared based on Akaike's Information Criterion (AIC [13]). A simulation study is performed in order to further investigate the ability of AIC to identify the underlying true model.

DATA

Measles incidence

In Germany, introduction of the measles vaccine had reduced the incidence of measles to a historical low of 0·2 cases/100 000 inhabitants in 2004 [3], before the disease re-emerged due to outbreaks in a few regions. We used measles surveillance data from Germany for the years 2005–2007, which contain weekly counts of cases for all ages in all 16 federal states reported to the RKI [11]. Figure 1 shows the notified measles cases in the years 2005–2007 for six selected federal states to illustrate the different incidence patterns. Large outbreaks occurred in Hesse and Bavaria in 2005 [3], in North Rhine-Westphalia in 2006 [4] and in North Rhine-Westphalia and Bavaria in 2007 [5]. The majority of cases ($\sim$80%) occurred in children and adolescents. About 12% occurred in infants aged $<$2 years. This pattern was very similar in all three years considered. A brief summary of the number of reported cases in each state is shown in Table 1 together with population numbers at 31 December 2006 obtained from the Federal Statistical Office of Germany [14].

Measles-mumps-rubella (MMR) vaccination

Coverage levels of the combined MMR vaccine were derived from vaccination cards presented at medical examinations, which are conducted by local health authorities at school entry [12]. Records include information about receipt of the first and second doses of MMR, but no information about dates or age of the child at vaccination. Age at school entry ranges between states from 4 to 7 years [15], therefore the information collected typically refers to vaccinations received 3–5 years previously [16].

The estimated coverage data do not include any information from children who did not present a vaccination card on the day of the medical examination (5–13% of children attending the school entry examination in different states). This is likely to overestimate true coverage, because the vaccination status of children with vaccination cards is generally more complete than in those without a card [4, 17]. However, there are no national data about the degree of overestimation. We made an assumption, which was used in a previous German study [18], that for each dose, the percentage of children without a vaccination card, 'non-card holders' was half that of 'card holders'. We applied this adjustment to all

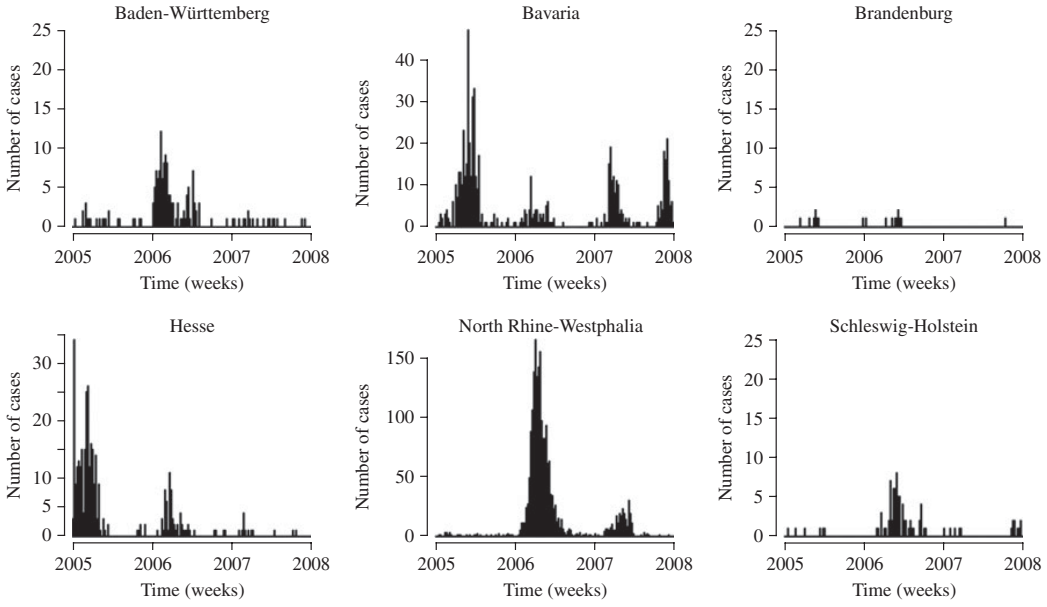


Fig. 1. Number of weekly measles cases in selected German federal states for the years 2005–2007. Note that the y-axis is not the same for all states.

analyses and conducted a sensitivity analysis to examine the robustness of the assumption.

Coverage levels for both the first and the second dose were higher in the new, re-established states in East Germany (Brandenburg, Mecklenburg-Western Pomerania, Saxony, Saxony-Anhalt, Thuringia) than in West Germany (Table 1). This might reflect continuing adherence to different childhood vaccination policies before re-unification [15, 19]. Immunization is voluntary in Germany now, but it was mandatory in the former German Democratic Republic.

METHODS

To investigate a possible association between the occurrence of measles epidemics and MMR vaccination coverage, we first examined the correlation between the number of observed cases in a region and region-specific vaccination coverage. One possibility is to apply the variance-stabilizing transformation for Poisson counts [20], i.e. taking the square root of cases, before estimating the empirical correlation coefficient which might improve the goodness of the corresponding confidence intervals. An alternative approach, based on a Poisson regression model [21, 22], assumes that the sum of cases in region i, aggregated over all three years, has mean

$$\mu_i = \exp(\alpha + \beta x_i), \tag{1}$$

where x_i denotes the coverage in state i. For example, to adjust for regionally varying population numbers, the right hand side of equation (1) can be multiplied by an offset n_i. Conclusions about the effect β of the covariate x_i in equation (1) remain the same when considering the weekly number of cases instead of the sum of cases, assuming that the weekly counts are independent. However, a multivariate time-series analysis of counts is able to incorporate autocorrelation and provides many more possibilities compared to the analysis of temporally aggregated data.

In the following, $y_{i,t}$ denotes the number of cases of a specific disease in a defined geographical region $i = 1, \ldots, I$ at time $t = 1, \ldots, T$. A fundamental assumption of a Poisson regression model is that the response variables $y_{i,t}$ are independent given the covariates. Thus the above model is not suited for the analysis of the measles data as the weekly counts are clearly dependent. Regular temporal dependence can easily be accounted for by including covariates for long-term or seasonal trends in the model. For instance, seasonal variation can be modelled parametrically using a superposition of harmonic waves [10, 23] or non-parametrically [9, 24]. However, such

Table 1. *Measles cases and estimated vaccination coverage in the 16 federal states of Germany*

State	Population	Measles cases		Coverage (%)		Presented cards (%)
		Max.	Sum	1st dose	2nd dose	
Baden-Württemberg (BW)	10 738 753	12	162	93·7	78·7	92·1
Bavaria (BY)	12 492 658	47	606	91·7	75·7	93·4
Bremen (HB)	663 979	1	4	94·6	76·9	86·9
Hamburg (HH)	1 754 182	3	29	93·9	84·0	91·7
Hesse (HE)	6 075 359	34	336	94·8	81·2	92·4
Lower Saxony (NI)	7 982 685	12	144	95·4	81·6	91·2
North Rhine-Westphalia (NW)	18 028 745	165	2036	95·2	81·6	88·5
Rhineland-Palatinate (RP)	4 052 860	9	85	94·9	80·8	91·4
Saarland (SL)	1 043 167	0	0	95·2	85·6	91·1
Schleswig-Holstein (SH)	2 834 254	8	89	94·7	83·6	89·8
Berlin (BE)	3 404 037	8	104	93·8	83·6	91·9
Brandenburg (BB)	2 547 772	2	18	97·1	89·8	93·5
Mecklenburg-Western Pomerania (MV)	1 693 754	1	4	97·5	91·6	92·1
Saxony (SN)	4 249 774	2	18	97·3	85·0	93·9
Saxony-Anhalt (ST)	2 441 787	2	12	97·7	89·8	92·6
Thuringia (TH)	2 311 140	3	8	97·4	88·3	94·6

Population estimated at 31 December 2006; maximum and total number of weekly measles cases from week 1, 2005 to week 52, 2007; coverage at school entry for the first and second dose of MMR vaccine in 2006 estimated from children presenting vaccination cards at school entry examinations; percentage of children with a vaccination card.

a model may still not adequately capture occasional outbreaks typical for infectious diseases.

A natural way to incorporate temporal dependence beyond seasonal variation is to consider the number of past cases as additional explanatory variables in the model. Held *et al.* [10] suggest a Poisson regression model with an identity link, where the (conditional) mean $\mu_{i,t}$ of $y_{i,t}$ is additively decomposed into two parts

$$\mu_{i,t} = \lambda y_{i,t-1} + \nu_{i,t}. \quad (2)$$

The first part with conditional rate $\lambda y_{i,t-1}$ is called the 'epidemic' component and the second part with rate $\nu_{i,t}$ the 'endemic' component. The former component captures occasional (epidemic) outbreaks whereas the latter describes regular (endemic) patterns.

To include region-specific covariate information, we allow the autoregressive parameter λ in equation (2) to vary across regions, i.e. we switch notation from λ to λ_i and model λ_i as a function of these covariates. Furthermore, covariates can also be considered in the other component $\nu_{i,t}$. Note that the conditional mean $\mu_{i,t}$ needs to be non-negative. This can be ensured by modelling both λ_i and $\nu_{i,t}$ on a log-scale.

Our first model (type A) assumes that the coverage levels in all states, x_i, enter into the epidemic component and the model is given by

$$\log(\lambda_i) = \beta_0 + \beta_1 x_i, \quad (3)$$

$$\log(\nu_{i,t}) = \alpha_0 + \{\gamma \sin(2\pi t/f) + \delta \cos(2\pi t/f)\} + \log(n_i), \quad (4)$$

where β_0 is an intercept and β_1 quantifies the influence of vaccination coverage. The parameter α_0 denotes the intercept of the endemic component and the offset $\log(n_i)$ represents population fractions, computed from Table 1. The terms in curly brackets in equation (4) are used to model seasonal variation. The number of data points per season is denoted by f. For instance, for a season of 1 year and weekly data $f = 52$. For ease of interpretation, the seasonal terms can be written equivalently as a sine wave with amplitude A describing the magnitude, and phase difference φ describing the onset of the seasonal pattern [23]. In the second model, the term $\beta_1 x_i$ is omitted in equation (3) and the coverage levels x_i are included instead in the endemic component with coefficient α_1. Altogether, the model (type B) is given by

$$\log(\lambda_i) = \beta_0, \quad (5)$$

$$\log(\nu_{i,t}) = \alpha_0 + \alpha_1 x_i + \{\gamma \sin(2\pi t/f) + \delta \cos(2\pi t/f)\} + \log(n_i). \quad (6)$$

Table 2. *Estimated Pearson's correlation coefficient, r, with 95% confidence intervals*

	Adjusted vaccination coverage			
	1st dose		2nd dose	
	r	95% CI	*r*	95% CI
Sum of cases	−0·34	−0·71 to 0·19	−0·34	−0·72 to 0·19
Square root of sum of cases	−0·44	−0·77 to 0·07	−0·48	−0·79 to 0·02

To investigate the impact of the explanatory variable, we also consider a model of type C, given by equations (4) and (5), where no covariate is included. Additionally, a standard log-linear Poisson regression model without the autoregressive component is fitted (model D).

For the model of type A we use the log proportion of unvaccinated school starters as explanatory variable x_i in equation (3) in accordance with the mass action principle [25]. This principle assumes that the rate of disease spread is proportional to the product of the density of susceptibles (unvaccinated school starters) multiplied by the density of infected individuals (reported cases). Taking the logarithm of the proportion of unvaccinated school starters produces the multiplicative relation (model A_0). Similarly, the log proportion of all school starters who received at most one dose of MMR vaccine is used as an explanatory variable. We used the same covariates in the model of type B.

Maximum likelihood (ML) estimates of parameters and standard errors (S.E.) are obtained by numerically maximizing the respective Poisson log-likelihood. Standard software for linear Poisson regression cannot be used because of the nonlinearity of the parameters. Therefore, the quasi-Newton BFGS method implemented in the R [26] function optim is used for optimization. The fitting procedure and the measles data are integrated in the R package surveillance ([27]; http://surveillance.r-forge.r-project.org). Note that models involving more than one covariate, time-varying covariates or additional seasonal terms at higher frequencies [28] can also be fitted with this function in surveillance.

The models investigated in the Results section are compared based on the model choice AIC criterion. We were particularly interested in the ability of AIC to distinguish between the model types A and B. In order to investigate this we conducted a simulation study (see Appendix).

RESULTS

The sum of cases over the years 2005–2007 in each state is negatively correlated with coverage for both the first and second dose of MMR vaccine (Table 2). Absolute correlation increases slightly when taking the square root of cases. However, the statistical evidence for correlation is weak, since the upper 95% confidence limits are always positive.

We describe here an analysis of the multivariate time-series of counts to further investigate the measles incidence patterns. The generation time [25] for measles, i.e. the average time between the onset of symptoms in one case and the onset of symptoms in a second case directly infected by the first, is about 10 days [25, 29]. We therefore aggregate measles cases in successive bi-weekly periods to better reflect this characteristic time-scale [30, 31]. AIC is used as a model choice criterion. The simulation study, discussed in detail in the Appendix, showed that this criterion is suitable for the comparison of the different model formulations.

The results of the analysis of the bi-weekly aggregated measles data are summarized in Table 3. All considered models contain an overall intercept α_0, a seasonal term and population fractions n_i as offset. The last two models in the table contain no covariates. When including only an intercept in the epidemic component (model C), the fit improves substantially compared to a model without autoregression (model D). The ML estimate of $\lambda = \exp(\beta_0)$ is quite high, $\hat{\lambda} = 0{\cdot}85$ (S.E. $= 0{\cdot}02$), which indicates a strong dependence on the number of counts at the previous time point after adjustment for seasonal effects. Consequently, the use of a Poisson regression model (without autoregression) seems inappropriate for these data. Indeed, the series of deviance residuals obtained from model D showed considerable autocorrelation compared with model C, which showed almost no autocorrelation.

Table 3. *Analysis of bi-weekly aggregated measles data*

				Epidemic component		Endemic component			
Model	log(L)	p	AIC	β_0 (S.E.)	β_1 (S.E.)	α_0 (S.E.)	α_1 (S.E.)	A (S.E.)	φ (S.E.)
Log proportion of unvaccinated school starters									
A_0	−1778·1	5	3566·1	3·01 (0·52)	1·38 (0·23)	1·78 (0·06)	—	0·66 (0·08)	−0·10 (0·12)
B_0	−1783·4	5	3576·8	−0·17 (0·02)	—	5·43 (0·69)	1·52 (0·29)	0·73 (0·09)	−0·10 (0·39)
Log proportion of school starters who received at most 1 dose of MMR vaccine									
A_1	−1787·1	5	3584·1	1·34 (0·31)	1·02 (0·21)	1·76 (0·06)	—	0·65 (0·08)	−0·08 (0·13)
B_1	−1790·7	5	3591·4	−0·17 (0·02)	—	3·59 (0·45)	1·17 (0·29)	0·71 (0·09)	−0·09 (0·41)
No covariates									
C	−1799·4	4	3606·8	−0·16 (0·02)	—	1·76 (0·06)	—	0·66 (0·08)	−0·06 (0·12)
D	−5213·9	3	10433·8	—	—	3·25 (0·03)	—	1·65 (0·04)	−0·52 (0·02)

The log-likelihood is denoted by log(L); p is the number of parameters and Akaike's Information Criterion (AIC) $= -2\log(L) + 2p$; lower AIC values indicate better fit. The parameters β_0 and α_0 denote intercepts; β_1 and α_1 denote the effect of the covariate; A and φ denote the amplitude and onset of the seasonal pattern. The standard error is denoted by S.E.

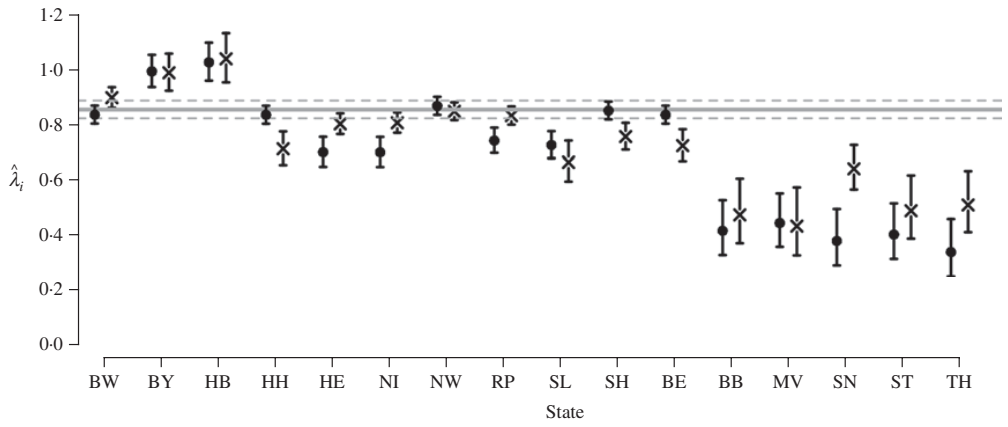


Fig. 2. Estimated autoregressive parameters $\hat{\lambda}_i$ and corresponding 95% confidence intervals for models A_0 (●) and A_1 (×). For comparison, the horizontal line denotes the estimated parameter $\hat{\lambda}$ for model C without covariates with the dashed lines representing the corresponding 95% confidence intervals. For definition of state abbreviations see Table 1.

In the next step, we investigated the impact of the inclusion of vaccination coverage in either the epidemic or endemic component compared to model C. Inclusion of the log proportion of unvaccinated school starters in the epidemic component (model A_0) leads to a considerably better fit.

The effect of the covariate β_1 in model A_0 is clearly significant ($P < 0{\cdot}0001$). Note that the estimated coefficients in the endemic component remain similar as in model C while the autoregressive parameter now varies across states. Inclusion of the covariate into the endemic component (model B_0) also improves the fit compared to model C but is worse compared to model A_0 according to AIC.

The above conclusions also hold when including the log proportion of school starters with at most one dose of MMR vaccine (models A_1, B_1). However, the model fit is considerably worse in terms of AIC. All results in Table 3 are based on the assumption that the coverage levels of the non-card holders are half those of card holders (adjustment factor 0·5). We tried several adjustment factors to investigate the robustness of our results. The ranking of the models according to AIC does not change for an adjustment factor $< 0{\cdot}6$. With regard to AIC an adjustment factor of 0·2 yields the best fit.

Figure 2 shows the estimated parameters λ_i and corresponding 95% confidence intervals for models

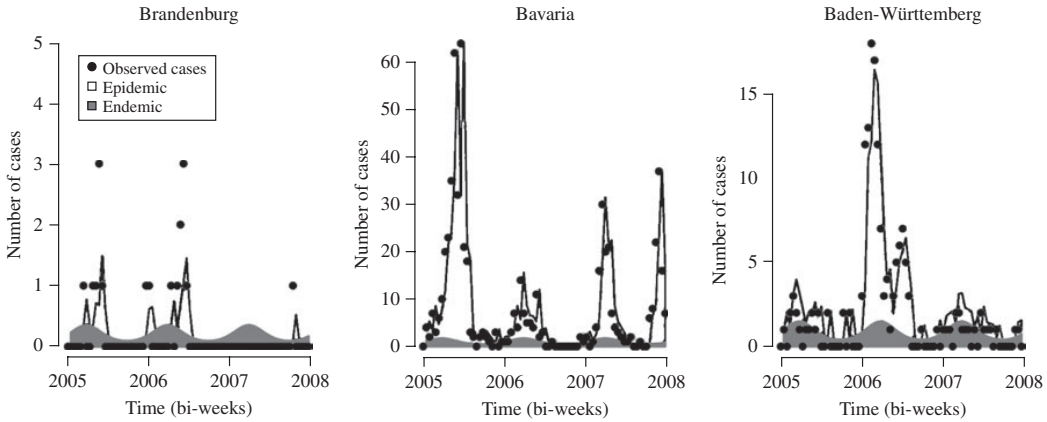


Fig. 3. Fitted mean for model A_0, which includes the log proportion of unvaccinated school starters as covariate in the epidemic component, in selected states.

A_0 and A_1 for each state. There is considerable heterogeneity across states. The ML estimates for the five states in East Germany are markedly lower than estimates for the remaining states. Vaccination coverage is considerably higher in these states. Note that model A_0 which includes the log proportion of unvaccinated school starters in the epidemic component performs better in terms of AIC than a model with the original (untransformed) proportion.

The analysis of the multivariate time-series of measles surveillance counts showed that there is an association between vaccination coverage and the occurrence and size of measles epidemics within states, with model A_0 fitting best. Figure 3 shows the fitted number of cases, decomposed into endemic and epidemic components, for this model in three of the states shown in Figure 1 for illustrative purposes. The estimated mean is clearly dominated by the epidemic component.

DISCUSSION

We observed a significant association between estimated vaccination coverage at school entry and the overall incidence of measles in the federal states of Germany (Table 3). The inclusion of the log proportion of unvaccinated school starters in the epidemic component of the model is the most suitable formulation to describe the occurrence and size of measles epidemics. This is plausible since the proportion of unvaccinated school starters acts as a proxy for the population of susceptibles, and the number of cases at a future time point depends on the number of infectious cases in the present as well as on the number of individuals susceptible to infection.

A strength of the proposed model is the decomposition of the disease incidence into an endemic and an epidemic component. Compared to a standard log-linear Poisson regression model our formulation is able to account for occasional outbreaks by including an autoregressive component. This is particularly important for the analysis of highly infectious diseases such as measles. In addition, information about vaccination coverage was included to cope with regional heterogeneity.

There are some limitations to this study. The RKI also provides estimates of vaccination coverage at school entry for children aged 4–7 for the years 2005 and 2007. However, the measles data comprise cases of all ages. Thus, changes in age-specific vaccination coverage may lead to shifts in the age distribution of the number of cases, but it will be impossible to discern such shifts from age-aggregated surveillance data. In addition, there is uncertainty about the true vaccination status, when obtained from school entry examinations. Hence small changes in coverage levels in successive years are not expected to be particularly meaningful. Therefore, we used only data for 2006 as an approximate measure of the overall immunization status in each state in all age groups.

We were aware that vaccination coverage was probably overestimated because vaccination uptake in school starters who presented vaccination cards is assumed to be higher [12]. Roughly 10% of school starters did not present vaccination cards and coverage for them is unknown. To assess the sensitivity of

the assumed coverage for those without cards (0·5 times that of card holders) we considered values ranging from the same coverage as children who presented cards (corresponding to 1) to all children who did not present cards being unvaccinated (corresponding to 0). In terms of AIC, model B where the covariate is included in the endemic component is not very sensitive with regard to the assumed coverage. In contrast, the AIC for model A where the covariate is included in the epidemic component changes considerably. When coverage for non-card holders is $>0{\cdot}6$ times that of card holders, model B is preferred.

Wichmann *et al*. [4] investigated a local outbreak in a school in Duisburg (North Rhine-Westphalia) in 2006. They estimated that receipt of one dose of MMR in the 22% without cards was 75% (significantly lower than the coverage of 95% in students with vaccination cards). This corresponds to a coverage level for non-card holders around 0·8 times that of card holders. However, this investigation involved only one school and no information about uncertainty around the estimated 75% coverage was given. The results are probably not generalizable to data at state level in this study. According to AIC, the measles data in our study are best described assuming coverage in non-card holders of 0·2 times that of card holders and using a model in which the proportion of unvaccinated school starters is incorporated in the epidemic component of the model.

To investigate the ability of AIC to identify the correct type of the model, we conducted a simulation study (Appendix). We used a simple model, comparable to the model of type A, where vaccination coverage influences the epidemic component. The simulation study showed that AIC identifies the true underlying model as long as the influence of vaccination coverage is strong or non-existent.

The proposed model approach allows us to consider infectious disease counts with several time-varying covariates. If quarterly, age-specific vaccination coverage was available, it could also be investigated whether vaccination-related trends in age-specific incidence [32] are observable using such notification data. Another interesting aspect would be to investigate the behaviour of the model where vaccination coverage is simultaneously included as an explanatory variable in both components. In this case, attention should be paid to potential issues related to multicollinearity or identifiability of parameter estimates.

Table 4. *Population sizes* (N_i) *and corresponding vaccination coverage levels* (x_i) *used in the simulation study*

Region	State	N_i	n_i	x_i	$\log(1-x_i)$
1	Bavaria	12 492 658	0·44	0·90	−2·30
2	Lower Saxony	7 982 685	0·28	0·85	−1·90
3	Saxony	4 249 774	0·15	0·85	−1·90
4	Berlin	3 404 037	0·12	0·80	−1·61

The states used in the simulation study were selected at random. The population fraction is denoted by n_i.

Table 5. *Models for the simulation analysis*

Model	Epidemic component $\log(\lambda_i)=\beta_0+$	Endemic component $\log(\nu_i)=\log(n_i)+\alpha_0+$
A	$\beta_1 \log(1-x_i)$	—
B	—	$\alpha_1 \log(1-x_i)$
C	—	—

In order to apply the proposed model to data at a finer spatial resolution we would need more detailed information about vaccination coverage because there are great regional and local differences leading to immunization gaps [6, 15]. For example, coverage levels for one dose of MMR vaccine ranged from 77·5% to 98% in the 77 health districts of Bavaria at school entry examinations 2005/2006 [33]. At a finer spatial resolution, it might also be necessary to account for spatio-temporal dependence, e.g. due to commuting. This could be done by including the previous number of cases in adjacent regions in the epidemic component [10, 23].

Although the data on measles incidence and vaccination coverage have some limitations, clear associations were observed. The pattern observed in the reported measles cases for all ages is best described by including the log proportion of unvaccinated school starters as an explanatory variable in the autoregressive (epidemic) component of the model.

APPENDIX: Simulation study

We investigated whether AIC identifies the correct structure of the model with a simulation study. Multivariate time-series of length $T=156$ (3 years of weekly data) were simulated based on a model where the number of cases $y_{i,t}$ in region i at time t is influenced by vaccination coverage as a covariate. Each

Table 6. *Results for the simulation study*

	Fixed parameters			Average number of cases				AIC % of model			
Sim	c	$\bar{\lambda}$	β_1	Reg 1	Reg 2	Reg 3	Reg 4	A	B	C	True model
1	10^{-4}	0·5	0	3813	2296	1201	1061	10·1	8·8	81·1	C
2	10^{-4}	0·5	0·1	3587	2613	1332	1014	26·1	19·8	54·1	A
3	10^{-4}	0·5	0·5	3232	2294	1259	1246	81·9	18·1	0·0	A
4	10^{-4}	0·8	0	4004	2396	1318	660	10·9	10·6	78·5	C
5	10^{-4}	0·8	0·1	3139	2377	1080	997	48·3	22·0	29·7	A
6	10^{-4}	0·8	0·5	2111	2247	1214	2162	99·3	0·7	0·0	A
7	10^{-5}	0·5	0	364	223	129	96	12·6	12·4	75·0	C
8	10^{-5}	0·5	0·1	376	249	141	98	14·1	11·6	74·3	A
9	10^{-5}	0·5	0·5	305	295	127	132	62·7	12·6	24·7	A
10	10^{-5}	0·8	0	333	256	206	107	13·7	13·8	72·5	C
11	10^{-5}	0·8	0·1	413	258	183	76	20·9	13·6	65·5	A
12	10^{-5}	0·8	0·5	234	146	146	359	87·1	3·9	9·0	A

AIC, Akaike's Information Criterion.
Parameter values are shown for the simulations (Sim), the mean number of cases for each region (Reg), and how often each model has the lowest AIC value (AIC % of model).

simulated dataset is analysed with different models and AIC is calculated.

We assumed that vaccination coverage influences the epidemic component, which also contains an intercept. The endemic component contains no seasonal terms, an overall intercept α_0 and population fractions n_i as offset. Four randomly selected regions are used where the population sizes N_i are selected from population data of Germany in 2006 and artificial vaccination coverage levels x_i are attached (see Table 4). The coverage levels x_i differ between the regions and have been transformed with $\log(1-x_i)$ as in the measles analysis. The simulation model corresponds to model A in Table 5 and is similar to model A_0 for the measles data (Table 3).

We chose different values for the yearly incidence c (10^{-4}, 10^{-5}) and the basic level of the epidemic component not influenced by covariates, $\bar{\lambda}$ (0·5, 0·8). Furthermore, we assumed that vaccination coverage has either no ($\beta_1=0$), a small ($\beta_1=0{\cdot}1$), or a strong ($\beta_1=0{\cdot}5$) influence. All combinations of these values give 12 different simulation scenarios. For each of these scenarios, 1000 datasets have been simulated. The incidence c and the population size N_i are used to calculate the mean number of cases for the first week $\mu_{i,1}$ for each region with $\mu_{i,1}=cN_i/52$. The parameter $\bar{\lambda}$ is used to calculate the intercept β_0 as a basic level

$$\beta_0=\log(\bar{\lambda})-\text{mean}(\beta_1\log(1-x_i)).$$

Next, the epidemic component λ_i is calculated as in model A (Table 5) and used for the simulation. The endemic component ν is calculated with the stationary mean equation [10]

$$\nu_i=\mu_{i,t}\frac{(1-\bar{\lambda})}{n_i}=\frac{cN_i}{52}\frac{(1-\bar{\lambda})\sum_i N_i}{N_i}$$
$$=\frac{c}{52}(1-\bar{\lambda})\sum_i N_i$$

and is the same for all regions. The cases $y_{i,t}$ are simulated for each region i and point in time t as follows:

$$y_{i,t}\sim Po\left(\frac{n_i\nu}{1-\lambda_i}\right) \qquad (t=1),$$
$$y_{i,t}\sim Po(\lambda_i y_{i,t-1}+n_i\nu) \quad (t=2,\ldots,T).$$

For the analysis of each simulated dataset three different models, listed in Table 5, have been considered. The models differ with regard to the influence of vaccination coverage: in the epidemic component, in the endemic component, or none. Note that the values of the covariates used in the analysis are the same as in the simulation.

The results of the analysis are shown in Table 6. In all simulations where there was no influence of vaccination coverage the true underlying model C resulted most frequently in the lowest AIC value (i.e. highest AIC %). When there was a small influence of vaccination coverage in the epidemic component, AIC in general preferred model C with no influence, followed by model A with influence in the epidemic component. When there was a strong influence, model A is clearly preferred. In summary, AIC identifies the true

underlying model as long as the influence of vaccination coverage is strong or non-existent.

ACKNOWLEDGEMENTS

We thank J. C. M. Heijne and N. Low for valuable comments and suggestions. Financial support by the Swiss National Science Foundation is gratefully acknowledged.

DECLARATION OF INTEREST

None.

REFERENCES

1. **Muscat M, *et al.*** Measles in Europe: an epidemiological assessment. *Lancet* 2009; **373**: 383–389.
2. **Wallinga J, Heijne JCM, Kretzschmar M.** A measles epidemic threshold in a highly vaccinated population. *PLoS Medicine* 2005; **2**: e316.
3. **Siedler A, *et al.*** Two outbreaks of measles in Germany 2005. *Eurosurveillance* 2006; **11**: 131–134.
4. **Wichmann O, *et al.*** Large measles outbreak at a German public school, 2006. *Pediatric Infectious Disease Journal* 2007; **26**: 782–786.
5. **Bernard H, *et al.*** An outbreak of measles in Lower Bavaria, Germany, January–June 2007. *Eurosurveillance* 2007; **12**: pii = 3278.
6. **Wichmann O, *et al.*** Further efforts needed to achieve measles elimination in Germany: results of an outbreak investigation. *Bulletin of the World Health Organization* 2009; **87**: 108–115.
7. **Cummings DAT, *et al.*** Improved measles surveillance in Cameroon reveals two major dynamic patterns of incidence. *International Journal of Infectious Diseases* 2006; **10**: 148–155.
8. **Frank C, *et al.*** Cattle density and Shiga toxin-producing Escherichia coli infection in Germany: increased risk for most but not all serogroups. *Vector-Borne and Zoonotic Diseases* 2008; **8**: 635–643.
9. **Hens N, *et al.*** Estimating the impact of vaccination using age-time-dependent incidence rates of hepatitis B. *Epidemiology and Infection* 2008; **136**: 341–351.
10. **Held L, Höhle M, Hofmann M.** A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling* 2005; **5**: 187–199.
11. **Robert Koch Institute.** SurvStat (http://www3.rki.de/SurvStat). Accessed 14 October 2009.
12. **Robert Koch Institute**. On vaccination coverage estimated at school entry examinations in Germany, 2006 [in German]. *Epidemiologisches Bulletin* 2008; **7**: 55–57.
13. **Lindsey JK, Jones B.** Choosing among generalized linear models applied to medical data. *Statistics in Medicine* 1998; **17**: 59–68.
14. **Statistisches Bundesamt.** 12411-0009 Projections of the resident population for German states on reference date (https://www-genesis.destatis.de/genesis/online/logon). Accessed 21 January 2009.
15. **Kalies H, *et al.*** Immunisation status of children in Germany: temporal trends and regional differences. *European Journal of Pediatrics* 2006; **165**: 30–36.
16. **Reiter S, Poethko-Müller C.** Current vaccination coverage and immunization gaps of children and adolescents in Germany [in German]. *Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz* 2009; **52**: 1037–1044.
17. **Poethko-Müller C, *et al.*** Vaccination coverage against measles in German-born and foreign-born children and identification of unvaccinated subgroups in Germany. *Vaccine* 2009; **27**: 2563–2569.
18. **Tischer A, Siedler A, Rasch G.** Surveillance of measles in Germany [in German]. *Gesundheitswesen* 2001; **63**: 703–709.
19. **Hellenbrand W, *et al.*** Progress toward measles elimination in Germany. *Journal of Infectious Diseases* 2003; **187** (Suppl. 1): S208–S216.
20. **Palmgren J.** Poisson distribution. In: Armitage P, Colton T, eds. *Encyclopaedia of Biostatistics*, 2nd edn. West Sussex: John Wiley and Sons, 2005, pp. 4109–4113.
21. **Kirkwood B, Sterne J.** *Essential Medical Statistics*, 2nd edn, 2003. Malden: Wiley.
22. **Kuhn L, Davidson LL, Durkin MS.** Use of Poisson regression and time series analysis for detecting changes over time in rates of child injury following a prevention program. *American Journal of Epidemiology* 1994; **140**: 943–955.
23. **Paul M, Held L, Toschke M.** Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine* 2008; **27**: 6250–6267.
24. **Knorr-Held L, Richardson S.** A hierarchical model for space-time surveillance data on meningococcal disease incidence. *Journal of the Royal Statistical Society Series C* 2003; **52**: 169–183.
25. **Anderson RM, May RM.** *Infectious Diseases of Humans: Dynamics and Control*. Oxford: Oxford University Press, 1991.
26. **R Development Core Team.** *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2009.
27. **Höhle M.** Surveillance: an R package for the monitoring of infectious diseases. *Computational Statistics* 2007; **22**: 571–582.
28. **Diggle PJ.** *Time Series. A Biostatistical Introduction*. New York: Oxford University Press, 1990.
29. **Fine PEM, Clarkson JA.** Measles in England and Wales – I: An analysis of factors underlying seasonal patterns. *International Journal of Epidemiology* 1982; **11**: 5–14.
30. **Cliff AD, Haggett P.** Statistical modelling of measles and influenza outbreaks. *Statistical Methods in Medical Research* 1993; **2**: 43–73.
31. **Finkenstädt BF, Grenfell BT.** Time series modelling of childhood diseases: A dynamical systems approach.

Journal of the Royal Statistical Society: Series C 2000; **49**: 187–205.

32. **Fine PEM, Clarkson JA.** Measles in England and Wales – II: The impact of the measles vaccination program on the distribution of immunity in the population. *International Journal of Epidemiology* 1982; **11**: 15–25.
33. **Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit.** MMR vaccination coverage of children starting school in Bavaria in 2006/07 at a regional level (http://www.lgl.bayern.de/gesundheit/gesundheitsindikatoren/themenfeld07/indikator0714.htm). Accessed 10 June 2009.

APPENDIX I

Incorporating random effects in a likelihood-based model for multivariate infectious disease counts

Michaela Paul & Leonhard Held

Proceedings of the 16^{th} *European Young Statisticians meeting*,
Bucharest, Romania, 2009

Incorporating random effects in a likelihood-based model for multivariate infectious disease counts

Michaela Paul, Leonhard Held

Biostatistics Unit, Institute of Social and Preventive Medicine, University of Zurich
Hirschengraben 84, 8001 Zurich, Switzerland

Abstract

In this paper we discuss a likelihood-based framework for analysing time series of infectious disease surveillance counts typically observed in several geographical areas. The model is extended to account for heterogeneous incidence levels or varying transmission of a pathogen across regions. This is done by introducing region-specific random effects in the predictor. Inference is based on variants of methodology for mixed models. For illustration, the model is applied to data on meningococcal disease in France. The predictive properties are investigated by means of one-step-ahead predictions and proper scoring rules.

Keywords: Multivariate time series of counts, infectious diseases, random effects, proper scoring rules.

1. Introduction

Surveillance data on infectious diseases usually consist of counts of new infections observed in defined geographical areas at regular time intervals. A simple modelling approach for such data is log-linear Poisson regression which accounts for temporal trends and seasonal variation in the mean. However, such a model can only describe regular patterns and cannot adequately capture occasional outbreaks typical for infectious diseases. Hence, a natural way to incorporate temporal dependence beyond seasonal variation is to consider the number of past cases as additional explanatory variables in the model.

Paul et al. (2008) [7] use a Poisson regression model with identity link where the disease incidence is divided into two additive components. The first component represents an autoregression on past counts, thus allowing for temporal dependence. The autoregressive parameter can either be the same or different across regions if the transmission of the pathogen differs across regions e.g. due to heterogeneous immunisation levels or other unknown factors. Similarly, the second component can account for varying incidence levels.

However, in the case of a large number of regions this approach of using region-specific (fixed) parameters can be problematic because some regions may contain only little information about the parameters leading to identifiability problems. Instead we will introduce random effects to deal with heterogeneity in some of the model coefficients. Inference is based on mixed model methodology [1, 4].

In this paper, we investigate the performance of the model including random effects. The predictive properties are investigated by means of one-step-ahead predictions and proper scoring rules. For illustration, the model is applied to meningococcal disease in France.

2. Modelling approach

2.1. Model

Denote $y_{i,t}$ the number of cases in region $i = 1, \ldots, I$ at time $t = 1, \ldots, T$. The counts are assumed to be Poisson distributed, $y_{i,t}|y_{i,t-1} \sim \text{Po}(\mu_{i,t})$, with conditional mean

$$\mu_{i,t} = \lambda_i y_{i,t-1} + \nu_{i,t}\,, \quad \lambda_i,\, \nu_{i,t} > 0. \tag{1}$$

As in Paul et al. (2008) [7] the first additive component of the conditional mean with rate $\lambda_i y_{i,t-1}$ is called the 'epidemic' component and the second component with rate $\nu_{i,t}$ is called the 'endemic' component. The former should capture occasional outbreaks whereas the latter should describe long-term trends and regular seasonal patterns.

The unknown coefficients $\log(\lambda_i)$ and $\log(\nu_{i,t})$ in (1) are decomposed additively

$$\log(\lambda_i) = \beta_\lambda + b_{\lambda,i} \tag{2}$$

$$\log(\nu_{i,t}) = \beta_\nu + b_{\nu,i} + \tau t + \sum_{s=1}^{S} \big(\gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t)\big) + \log(n_{i,t}) \tag{3}$$

with Fourier frequencies $\omega_s = 2\pi s/12$ for monthly data and offset $n_{i,t}$, e.g. yearly varying population numbers.

To account for heterogeneity across units, a random intercept is included in both components (2) and (3) of the mean. The parameters $\boldsymbol{b}_\lambda = (b_{\lambda,1}, \ldots, b_{\lambda,I})^T$ and $\boldsymbol{b}_\nu = (b_{\nu,1}, \ldots, b_{\nu,I})^T$ are assumed to be normally distributed with mean vector $\mathbf{0}$ and covariance matrix

$$\boldsymbol{\Sigma} = \text{diag}(\sigma^2_\lambda \mathbf{I}_I, \sigma^2_\nu \mathbf{I}_I), \tag{4}$$

where $\mathbf{I}_I$ is the $I \times I$ identity matrix. Hence, the vectors of the fixed (unpenalised) and random (penalised) coefficients are given by the vectors $\boldsymbol{\beta} = (\beta_\lambda,\, \beta_\nu, \tau, \gamma_1, \ldots, \gamma_S, \delta_1, \ldots, \delta_S)^T$ and $\boldsymbol{b} = (\boldsymbol{b}_\lambda^T, \boldsymbol{b}_\nu^T)^T$.

2.2. Estimation

The estimation procedure is based on variants of the mixed model methodology proposed by Breslow and Clayton (1993) [1] and developed further for survival data in Kneib and Fahrmeir (2007) [4]. It can be viewed as penalised likelihood inference from a frequentist perspective or as an empirical Bayes approach from a Bayesian perspective. While flat and Gaussian priors are specified for the fixed and random effects, respectively, the variance components are treated as fixed and are estimated in advance from the data by maximising the (approximate) marginal likelihood. This leads to the following alternating algorithm for the estimation of all parameters.

1. Estimation of regression parameters for given variance components

Inference for the regression parameters, given $\sigma^2_\lambda, \sigma^2_\nu$, is based on the penalised log-likelihood

$$l_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{\Sigma}) = \sum_{i,t} y_{i,t} \log(\mu_{i,t}) - \mu_{i,t} - \sum_{r \in \{\lambda,\nu\}} \frac{1}{2\sigma^2_r} \boldsymbol{b}_r^T \boldsymbol{b}_r. \tag{5}$$

First and second derivatives with respect to the regression parameters can be easily derived and allow the computation of updated estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}$ given the variances via Newton-Raphson steps.

2. Estimation of variance components for given regression parameters

Estimates of the variance components are obtained by maximising the marginal likelihood

$$L_{\text{marg}}(\boldsymbol{\Sigma}) = \int \exp\left\{l_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\Sigma})\right\} d\boldsymbol{\beta} d\boldsymbol{b}.$$

The above integral cannot be solved analytically. Applying a Laplace approximation to the integrand and assuming that small changes in the variance parameters do hardly affect the estimates of the regression coefficients [4] results in an approximate marginal log-likelihood

$$l_{\text{marg}}(\boldsymbol{\Sigma}) \approx -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\boldsymbol{b}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{b} - \frac{1}{2}\log|\boldsymbol{F}_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{\Sigma})| \tag{6}$$

where $|\cdot|$ denotes the determinant of a matrix and $\boldsymbol{F}_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{\Sigma})$ denotes the observed Fisher information matrix for the regression parameters based on the penalised likelihood (5).

First and second derivatives of (6) can be derived based on differentiation rules for matrices. The score function $\boldsymbol{s}_{\text{marg}}(\boldsymbol{\Sigma})$ has elements

$$s_r = \frac{\partial l_{\text{marg}}(\boldsymbol{\Sigma})}{\partial \sigma_r^2} = -\frac{I}{2\sigma_r^2} + \frac{1}{2\sigma_r^4}\boldsymbol{b}_r^T\boldsymbol{b}_r + \frac{1}{2\sigma_r^4}\operatorname{tr}(\boldsymbol{G}_{rr}) \tag{7}$$

where $\operatorname{tr}(\cdot)$ denotes the trace of a matrix and $\boldsymbol{G}_{rr}$ is the diagonal block of the inverse of the Fisher information matrix $\boldsymbol{F}_{\text{pen}}^{-1}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{\Sigma})$ corresponding to $\boldsymbol{b}_r$, $r \in \{\lambda, \nu\}$. The elements of the observed Fisher information $\boldsymbol{F}_{\text{marg}}(\boldsymbol{\Sigma})$ are given as

$$F_{qr} = -\frac{\partial^2 l_{\text{marg}}(\boldsymbol{\Sigma})}{\partial \sigma_q^2 \sigma_r^2} = -\frac{1}{2\sigma_q^4\sigma_r^4}\operatorname{tr}(\boldsymbol{G}_{qr}\boldsymbol{G}_{rq}) - \mathbb{1}(q = r)\left[\frac{I}{2\sigma_r^4} - \frac{1}{\sigma_r^6}\boldsymbol{b}_r^T\boldsymbol{b}_r - \frac{1}{\sigma_r^6}\operatorname{tr}(\boldsymbol{G}_{rr})\right], \tag{8}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. Updated variances can be obtained via Newton-Raphson steps using the score vector (7) and the Fisher information matrix (8).

3. Iterate steps 1. and 2. until convergence

2.3. Predictive model assessment

Since the validation of models based on one-step-ahead predictions is particularly suited for the analysis of time series, we will assess the predictive properties of the model based on proper scoring rules [3] as an alternative to classical model choice criteria. Scoring rules evaluate a model based on the prediction and the actual observed value. They are negatively oriented penalties that are to be minimised.

Perhaps the most widely known scoring rule is the logarithmic score

$$\text{logS}(P, y) = -\log(\text{P}(Y = y)),$$

Table 1: Analysis of meningococcal disease in France. The parameter $A = \sqrt{\beta_\gamma^2 + \beta_\delta^2}$ denotes the amplitude and $\varphi = \arctan(\beta_\delta/\beta_\gamma)$ the phase difference of the seasonal component in (3) (see [7]), $\star$ indicates fixed but unit-specific intercepts. Average scores are based on 94×84 one-step-ahead predictions with p-values based on permutation tests for paired observations (9999 permutations) with model 3 as the reference model.

model	$\exp(\hat{\beta}_\lambda)$ (se)	$\hat{\beta}_\nu$ (se)	$\hat{\beta}_\tau$ (se)	$\hat{A}$ (se)	$\hat{\varphi}$ (se)	$\hat{\sigma}^2_\lambda$	$\hat{\sigma}^2_\nu$	l_{pen}	$\overline{\text{logS}}$ (p-value)	$\overline{\text{RPS}}$ (p-value)
1	–	0.168 (0.027)	−0.0020 (0.0003)	0.345 (0.020)	1.037 (0.018)	–	–	-10831	0.5970 (0.0001)	0.2026 (0.0001)
2	0.085 (0.009)	$\star$	−0.0019 (0.0003)	0.351 (0.022)	1.095 (0.019)	–	–	-10427	0.5867 (0.0012)	0.1977 (0.0080)
3	0.086 (0.009)	−0.028 (0.055)	−0.0019 (0.0003)	0.352 (0.029)	1.097 (0.019)	–	0.18	-10468	**0.5844**	**0.1975**
4	0.089 (0.010)	−0.049 (0.057)	−0.0019 (0.0003)	0.356 (0.022)	1.120 (0.019)	0.41	0.19	-10447	0.5852 (0.1027)	0.1978 (0.2019)

where P is the predictive distribution and y denotes the count that materialises. Another scoring rule which is less sensitive to extreme events is the ranked probability score

$$\text{RPS}(P, y) = \sum_{k=0}^{\infty} \Big(\text{P}(Y \leq k) - \mathbb{1}(y \leq k) \Big)^2.$$

See Czado et al. (2009) [2] for a discussion of these scores.

Typically mean scores over a set of predictions are used to rank and compare different models informally or via tests. Here we will look at a series of one-step-ahead predictions for each considered model. Two models are compared with a permutation test for paired observations [6].

3. Meningococcal disease in France

In the following, we analyse the monthly incidence of meningococcal disease in the 94 départments of France excluding Corsica from 1985 to 1999. The data have been previously analysed in Knorr-Held and Richardson (2003) [5]. Table 1 summarises the obtained estimates for selected models. The validity of the models is assessed through average scores based on one-step-ahead predictions of the last five years. All models contain a linear trend, $S = 1$ seasonal component and include expected cases $e_{i,t}$, calculated by indirect age-sex standardisation [5], as offset in (3).

The estimates for the linear trend and the seasonal components are nearly identical for all models. Inclusion of some form of autoregression in the model clearly improves the fit and the predictive performance. Models 3 with random intercepts $\boldsymbol{b}_\nu$ performs significantly better with respect to both scores than models 1-2 with a single and region-specific fixed intercepts,

respectively. Although the variance σ_λ^2 is estimated fairly large in model 4, which indicates considerable heterogeneity in the autoregressive coefficients, the average scores are slightly larger than for model 3. However, these score differentials are not important according to the p-values of the permutation tests.

4. Discussion

In this paper we have discussed the incorporation of random effects in a likelihood-based model for the analysis of multivariate time series of infectious disease counts. If heterogeneity is present, the induced shrinkage of region-specific estimates towards the overall mean improves the predictive performance.

The model formulation can easily be adjusted for overdispersion by replacing the Poisson distribution with the negative binomial. The inclusion of an additional autoregression on the number of past cases in other regions to model dependencies between regions is also straightforward. See Paul et al. (2008) [7] for further information.

The algorithm assumes that the random effects are uncorrelated and priors are proper as is the case in this paper. However, this assumption is not too restrictive because the covariance matrix needs to be not necessarily the identity matrix [4]. For instance, instead of an i.i.d. Gaussian prior one could assume a spatial smoothness prior for the incidence levels $\boldsymbol{b}_\nu$ or let the seasonal parameters γ_1, δ_1 change smoothly over time.

Acknowledgements: This research is supported by the Swiss National Science Foundation.

REFERENCES

[1] Breslow, N.E. and Clayton, D.G. (1993), Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, 88, 9–25.

[2] Czado, C., Gneiting, T. and Held, L. (2009), Predictive model assessment for count data, *Biometrics*, in press, DOI:10.1111/j.1541-0420.2009.01191.x.

[3] Gneiting, T. and Raftery, A.E. (2007), Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association*, 102, 359–378.

[4] Kneib, T. and Fahrmeir, L. (2007), A mixed model approach for geoadditive hazard regression, *Scandinavian Journal of Statistics*, 34, 207–228.

[5] Knorr-Held, L. and Richardson, S. (2003), A hierarchical model for space-time surveillance data on meningococcal disease incidence, *Journal of the Royal Statistical Society: Series C*, 52, 169–183.

[6] Ludbrook, J. and Dudley, H. (1998), Why permutation tests are superior to t and F tests in biomedical research, *American Statistician*, 52, 127–132.

[7] Paul, M., Held, L. and Toschke, A.M. (2008), Multivariate modelling of infectious disease surveillance data, *Statistics in Medicine*, 27, 6250–6267.

APPENDIX II

Inference details

Inference details

In the following we discuss details of the estimation procedure for a general formulation of the model proposed in Papers I – IV. The model includes both (correlated) random effects and covariates. Maximum likelihood estimation in mixed effects models is numerically challenging, as the log-likelihood function involves a multidimensional integral. In most cases, a closed-form solution is not available. Several methods to compute the integral have been suggested in the literature, see e.g. Breslow and Clayton (1993); Pinheiro and Bates (1995); Wolfinger (1999); Rabe-Hesketh, Skrondal and Pickles (2002), and references therein. Approaches include numerical integration procedures such as adaptive Gaussian quadrature, and approximations to the log-likelihood such as Laplace approximations to avoid the integration.

In the following a penalized likelihood approach (e.g. Kneib and Fahrmeir, 2007; Breslow and Clayton, 1993) is adapted, as discussed in Paper III. Here, integration over the random effects is avoided at the cost of having to jointly maximize the log-likelihood with respect to both fixed and random effects. Variance estimates are obtained in a second step using an approximate marginal likelihood based on a first-order Laplace approximation (Tierney and Kadane, 1986). Higher-order approximations may be used to improve the accuracy of the approximation (Raudenbush, Yang and Yosef, 2000). Alternatively, the random effects could have been integrated out using numerical techniques or Monte Carlo approaches to obtain estimates of the fixed and variance effects. Such an approach might be more accurate in some situations. However, it is also computationally much more demanding.

To make this appendix self-contained, the model formulation is first presented based on the notation introduced in Paper III. Afterwards, log-likelihoods, score functions and Fisher information matrices required for the estimation procedure proposed in Paper III are derived.

1 Model formulation

Let y_{rt} denote the number of cases of a specific disease in 'unit' $r = 1, \ldots, R$ at time $t = 1, \ldots, T$ and $\boldsymbol{y}_t = (y_{1,t}, \ldots, y_{R,t})^\top$ the vector of counts at time t. We mainly consider spatially stratified counts where a unit corresponds to a certain region. However, a unit might just as well represent e.g. different age groups. In the following, we use the terms 'region' and 'unit' interchangeably.

The counts are assumed to be either Poisson distributed, $y_{rt}|\boldsymbol{y}_{t-1} \sim \text{Po}(\mu_{rt})$, or negative binomial distributed, $y_{rt}|\boldsymbol{y}_{t-1} \sim \text{NegBin}(\mu_{rt}, \psi)$ with conditional mean

$$\mu_{rt} = \lambda_{rt} y_{r,t-1} + \phi_{rt} \sum_{q \neq r} w_{qr} y_{q,t-1} + \nu_{rt}\,, \quad (\lambda_{rt}, \phi_{rt}, \nu_{rt} > 0)\,, \tag{1}$$

and conditional variance

$$\mathrm{v}_{rt} = \begin{cases} \mu_{rt} & \text{for Poisson} \\ \mu_{rt}(1 + \psi\mu_{rt})\,, \quad \psi > 0, & \text{for negative binomial.} \end{cases} \tag{2}$$

In the following, the three additive components of the mean μ_{rt} in (1) will be called the autoregressive, neighbor-driven, and endemic component, respectively. The unknown quantities

$\log(\lambda_{rt})$, $\log(\phi_{rt})$ and $\log(\nu_{rt})$ are decomposed additively

$$\log(\lambda_{rt}) = \alpha^{(\lambda)} + b_r^{(\lambda)} + \boldsymbol{x}_{rt}^{(\lambda)\top}\boldsymbol{\beta}^{(\lambda)} \tag{3}$$

$$\log(\phi_{rt}) = \alpha^{(\phi)} + b_r^{(\phi)} + \boldsymbol{x}_{rt}^{(\phi)\top}\boldsymbol{\beta}^{(\phi)} \tag{4}$$

$$\log(\nu_{rt}) = \alpha^{(\nu)} + b_r^{(\nu)} + \boldsymbol{x}_{rt}^{(\nu)\top}\boldsymbol{\beta}^{(\nu)} + \log(e_{rt}) \tag{5}$$

where $\alpha^{(\lambda)}$, $\alpha^{(\phi)}$, $\alpha^{(\nu)}$ are component-specific intercepts, $\boldsymbol{\beta}^{(\lambda)} = (\beta_1^{(\lambda)}, \ldots, \beta_I^{(\lambda)})^\top$, $\boldsymbol{\beta}^{(\phi)} = (\beta_1^{(\phi)}, \ldots, \beta_J^{(\phi)})^\top$, $\boldsymbol{\beta}^{(\nu)} = (\beta_1^{(\nu)}, \ldots, \beta_K^{(\nu)})^\top$ are vectors of unknown parameters, $\boldsymbol{x}_{rt}^{(\lambda)}$, $\boldsymbol{x}_{rt}^{(\phi)}$, $\boldsymbol{x}_{rt}^{(\nu)}$ are the corresponding covariate vectors which contain possibly region-specific and time-dependent covariate information, and $\log(e_{rt})$ is an offset. In particular, the covariate vector $\boldsymbol{x}_{rt}^{(\nu)}$ in (5) may contain a trend variable, or sine and cosine functions of time to account for seasonal variation (Serfling, 1963). For instance, for weekly data, a linear trend and two pairs of harmonic terms in the endemic component correspond to $\boldsymbol{x}_{rt}^{(\nu)} = (t, \sin(\omega_1 t), \cos(\omega_1 t), \sin(\omega_2 t), \cos(\omega_2 t))^\top$, with Fourier frequencies $\omega_s = 2\pi s/52$, $s = 1, 2$. All in all, the vector of fixed (unpenalized) effects is given by $\boldsymbol{\beta} = (\alpha^{(\lambda)}, \alpha^{(\phi)}, \alpha^{(\nu)}, \boldsymbol{\beta}^{(\lambda)\top}, \boldsymbol{\beta}^{(\phi)\top}, \boldsymbol{\beta}^{(\nu)\top})^\top$.

Random effects specification

The stacked vector $\boldsymbol{b} = (\boldsymbol{b}^{(\lambda)\top}, \boldsymbol{b}^{(\phi)\top}, \boldsymbol{b}^{(\nu)\top})^\top$ containing all random effects from the three components is assumed to follow a normal distribution with mean $\mathbf{0}$ and positive definite covariance matrix

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \text{blockdiag}(\sigma_\lambda^2 \mathbf{I}_R, \sigma_\phi^2 \mathbf{I}_R, \sigma_\nu^2 \mathbf{I}_R)\,, \tag{6}$$

where $\boldsymbol{\theta} = (\sigma_\lambda^2, \sigma_\phi^2, \sigma_\nu^2)^\top$ are unknown variance parameters and $\mathbf{I}_R$ denotes the $R \times R$ identity matrix.

The use of the identity matrix implies that all random effects are assumed to be uncorrelated. Although this assumption is convenient, it might be restrictive in some applications. To allow for correlations between different components, the covariance matrix (6) is replaced by

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Omega}(\boldsymbol{\theta}) \otimes \mathbf{I}_R\,, \tag{7}$$

where $\boldsymbol{\Omega}(\boldsymbol{\theta})$ is an unknown 3×3 covariance matrix and $\otimes$ denotes the Kronecker product. As the identity matrix $\mathbf{I}_R$ is positive definite and the Kronecker product of positive definite matrices is again positive definite, the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is positive definite if $\boldsymbol{\Omega}(\boldsymbol{\theta})$ is positive definite.

Note that with this formulation, the vector of random effects within a component, say $\boldsymbol{b}^{(\nu)}$, is uncorrelated across regions. The case of spatially correlated random effects will be covered later on. First, a suitable parameterization of $\boldsymbol{\Omega}(\boldsymbol{\theta})$ is discussed.

Parameterization of the random effects covariance matrix

Pinheiro and Bates (1996) compared several parameterizations for unstructured covariance matrices that ensure positive definiteness and allow the use of unconstrained optimization routines. The authors suggest a parameterization based on the Cholesky decomposition combined with spherical coordinates.

We choose to factorize the 3×3 covariance matrix $\boldsymbol{\Omega}(\boldsymbol{\theta})$ in terms of standard deviations and

correlations, so that

$$\boldsymbol{\Omega}(\boldsymbol{\theta}) = \mathbf{S}\mathbf{T}\mathbf{T}^\top\mathbf{S}\,,$$

where $\mathbf{S} = \text{diag}(\sigma_1, \sigma_2, \sigma_3)$ is a diagonal matrix with standard deviations and $\mathbf{T}$ is a lower triangular matrix with elements corresponding to the Cholesky decomposition of a 3×3 correlation matrix.

Such a parameterization ensures positive definiteness and is computationally simple and stable (Pinheiro and Bates, 1996). However, it lacks a direct interpretation of $\boldsymbol{\theta}$ in terms of the variances and correlations in $\boldsymbol{\Omega}(\boldsymbol{\theta})$. Therefore, we parameterize the matrix $\mathbf{T}$ in terms of spherical coordinates (Pinheiro and Bates, 1996)

$$\mathbf{T} = \begin{pmatrix} 1 & & \\ \cos(t_1) & \sin(t_1) & \\ \cos(t_2) & \cos(t_3)\sin(t_2) & \sin(t_3)\sin(t_2) \end{pmatrix}$$

(non-given entries are zero).

To ensure uniqueness of the spherical parameterization, it is required that $\sigma_i > 0$, $i = 1, 2, 3$, and $t_i \in (0, \pi)$, $i = 1, 2, 3$. This can be obtained by setting $s_i = \log(\sigma_i)$ and $r_i = \pi/2 - \arctan(t_i)$ (Rapisarda, Brigo and Mercurio, 2007). The matrix $\mathbf{T}$ can be simplified through basic trigonometric equalities. The final parameterization of the covariance matrix is given by

$$\boldsymbol{\Omega}(\boldsymbol{\theta}) = \begin{pmatrix} \exp(2s_1) & & \\ \frac{r_1\,\exp(s_2+s_1)}{\sqrt{r_1^2+1}} & \exp(2s_2) & \\ \frac{r_2\,\exp(s_3+s_1)}{\sqrt{r_2^2+1}} & \frac{\left(r_1\,r_2\,\sqrt{r_3^2+1}+r_3\right)\exp(s_3+s_2)}{\sqrt{r_1^2+1}\,\sqrt{r_2^2+1}\,\sqrt{r_3^2+1}} & \exp(2s_3) \end{pmatrix} \tag{8}$$

with $\boldsymbol{\theta} = (s_1, s_2, s_3, r_1, r_2, r_3)^\top$. For the sake of clarity, only the lower triangular of the symmetric matrix $\boldsymbol{\Omega}(\boldsymbol{\theta})$ is displayed.

The diagonal elements correspond to the variances of the random effects in the autoregressive (3), neighbor-driven (4), and endemic component (5). For instance, we have $\sigma_\lambda^2 = \exp(2s_1)$. Similarly, we obtain the correlation parameters, e.g. $\rho_{\lambda\nu} = r_2/\sqrt{r_2^2+1}$. In the case of uncorrelated random effects, the matrix $\mathbf{T}$ corresponds to a 3-dimensional identity matrix and $\boldsymbol{\Omega}(\boldsymbol{\theta})$ is a diagonal matrix with elements $\exp(2s_i)$, $i = 1, 2, 3$.

Spatially correlated (CAR) random effects

In hierarchical models for spatio-temporal data, it is often reasonable to assume spatially correlated random effects rather than independently and identically distributed (IID) ones. For instance, one might adopt a conditional autoregressive (CAR) model (Besag, York and Mollié, 1991; Rue and Held, 2005, Section 3.3.2) for the vector of random effects $\boldsymbol{b}^{(\nu)}$, say.

The CAR model assumes that effects in adjacent regions are more alike than effects in distant regions. The respective identity matrix in (6) is then replaced by an adjacency matrix $\mathbf{K}$ with

elements (Rue and Held, 2005, p. 102)

$$k_{rs} = \begin{cases} n_r & \text{if } r = s \\ -1 & \text{if } r \sim s \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where n_r denotes the number of neighbors of region r, and $r \sim s$ denotes that region r is a neighbor of (shares a common border with) region s. As the rows and columns of $\mathbf{K}$ sum to zero, $\mathbf{K}$ is not of full rank but of rank $g < R$, leading to an improper distribution. If all regions are connected, i.e. there are no 'islands', the rank deficiency equals $R - g = 1$. Otherwise, higher rank deficiencies are obtained. In this thesis, only connected regions are considered and thus, $\text{rank}(\mathbf{K}) = g = R - 1$.

The rank deficiency of $\mathbf{K}$ implies that its inverse, which is needed to obtain marginal likelihood estimates for the variance components, does not exist. Following Rue and Held (2005, p. 91) and Kneib and Fahrmeir (2007), the parameter vector $\boldsymbol{b}^{(\nu)}$ can be decomposed into a one-dimensional unpenalized (fixed) part, and a g–dimensional penalized part, which is IID Gaussian of the form

$$\boldsymbol{b}^{(\nu)} = (\mathbf{1}, \mathbf{Z}) \begin{pmatrix} \tilde{\alpha}^{(\nu)} \\ \tilde{\boldsymbol{b}}^{(\nu)} \end{pmatrix} = \tilde{\alpha}^{(\nu)} + \mathbf{Z}\tilde{\boldsymbol{b}}^{(\nu)} \,. \tag{10}$$

Here, $\mathbf{1}$ denotes the R-dimensional vector of ones, which forms a one-dimensional basis of the null space (Harville, 1997, p.139) of $\mathbf{K}$, and is orthogonal to $\mathbf{Z}$. The $R \times g$ matrix $\mathbf{Z}$ can be obtained from the spectral decomposition (Harville, 1997, p. 537) of $\mathbf{K} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top$, where the $g \times g$ diagonal matrix $\mathbf{D}$ contains the positive eigenvalues of $\mathbf{K}$ in descending order, and the $R \times g$ orthogonal matrix $\mathbf{Q}$ contains the corresponding eigenvectors. By setting $\mathbf{L} = \mathbf{Q}\mathbf{D}^{1/2}$, the matrix $\mathbf{Z}$ is then constructed by

$$\mathbf{Z} = \mathbf{L}(\mathbf{L}^\top \mathbf{L})^{-1}. \tag{11}$$

Such a decomposition (10) has properties (Kneib and Fahrmeir, 2007):

1. $\text{rank}(\,(\mathbf{1}, \mathbf{Z})\,) = R$
2. $\mathbf{1}^\top \mathbf{Z} = \mathbf{0}$
3. $\mathbf{1}^\top \mathbf{K} \mathbf{1} = \mathbf{0}$
4. $\mathbf{Z}^\top \mathbf{K} \mathbf{Z} = \mathbf{I}_g$

Thus, the composed matrix $(\mathbf{1}, \mathbf{Z})$ represents a one-to-one transformation of $\boldsymbol{b}^{(\nu)}$. The parameter $\tilde{\alpha}^{(\nu)}$ is unpenalized and the g–dimensional vector $\tilde{\boldsymbol{b}}^{(\nu)}$ is IID Gaussian, $\tilde{\boldsymbol{b}}^{(\nu)} \sim \mathcal{N}(\mathbf{0}, \sigma_\nu^2 \mathbf{I}_g)$.

The endemic component (5) can thus always be written as

$$\log(\nu_{rt}) = \alpha^{(\nu)} + \boldsymbol{z}_{rt}^{(\nu)\top} \boldsymbol{b}^{(\nu)} + \boldsymbol{x}_{rt}^{(\nu)\top} \boldsymbol{\beta}^{(\nu)} + \log(e_{rt}),$$

where $\alpha^{(\nu)}$ is a (fixed) intercept and

- assuming $\boldsymbol{b}^{(\nu)}$ follows a CAR model: the $(R-1) \times 1$ vector $\boldsymbol{z}_{rt}^{(\nu)}$ corresponds to the r-th row of the matrix $\mathbf{Z}$ in (11), and $\boldsymbol{b}^{(\nu)} \sim \mathcal{N}(\mathbf{0}, \sigma_\nu^2 \mathbf{I}_{R-1})$;
- assuming $\boldsymbol{b}^{(\nu)}$ are IID Gaussians: the $R \times 1$ vector $\boldsymbol{z}_{rt}^{(\nu)}$ corresponds to a R-dimensional unit vector whose r-th element equals 1 and the remaining $R - 1$ elements equal 0, and $\boldsymbol{b}^{(\nu)} \sim \mathcal{N}(\mathbf{0}, \sigma_\nu^2 \mathbf{I}_R)$.

To facilitate the writing down of derivatives in the following section, the vector of all fixed and random parameters may be divided into sub-groups of parameters as follows: Let $\boldsymbol{\xi}^{(\nu)} = (\alpha^{(\nu)}, \boldsymbol{\beta}^{(\nu)}, \boldsymbol{b}^{(\nu)})^\top$ contain all fixed and (possibly reparameterized) random parameters in the endemic component (5), and let $\boldsymbol{u}_{rt}^{(\nu)} = (1, \boldsymbol{x}_{rt}^{(\nu)\top}, \boldsymbol{z}_{rt}^{(\nu)\top})^\top$ denote the corresponding design vector in region r at time t. The vectors $\boldsymbol{\xi}^{(\lambda)}$, $\boldsymbol{u}_{rt}^{(\lambda)}$ and $\boldsymbol{\xi}^{(\phi)}$, $\boldsymbol{u}_{rt}^{(\phi)}$ are defined in analogy for components (3) and (4), respectively. The stacked vector $\boldsymbol{\xi} = (\boldsymbol{\xi}^{(\lambda)\top}, \boldsymbol{\xi}^{(\phi)\top}, \boldsymbol{\xi}^{(\nu)\top})^\top$ thus contains all fixed and random parameters except the overdispersion parameter ψ in (2). Moreover, let $\boldsymbol{\beta}$ contain all fixed parameters and, if applicable, the log overdispersion parameter $\tilde{\psi} = -\log(\psi)$. Note that in applications, each component (3)–(5) of the conditional mean may be omitted in parts or as a whole, and the parameter and design vectors are specified accordingly.

2 Inference

Parameter estimates are obtained via the alternating algorithm discussed in Paper III. First, random effects with associated improper distribution (CAR) are reparameterized to obtain random effects with proper distribution (IID). Then, the following two steps are iterated until convergence is reached:

1. Update the regression parameters given the current variance components through Newton steps.
2. Update variance components given current regression parameters through Newton steps.

Because the (penalized) log-likelihood is not necessarily log-concave, the standard Newton-Raphson (NR) algorithm requires initial values close to the optimum to ensure convergence. Far from the solution, the NR method may run into numerical problems as the Fisher information matrix may not be positive definite. Also, the algorithm might get stuck and not converge to the global maximum in the case of poor initial values. Such difficulties may be dealt with by adequate modifications leading to a globally convergent NR algorithm (see Dennis and Schnabel, 1996, Chapter 6).

In Paper II, ML estimates in a fixed-effects model are obtained using the BFGS method implemented in the `R` (R Development Core Team, 2010) function `optim`. In applications, a grid of initial values is used to find the global maximum. Should the optimization algorithm lead to different results, the estimates resulting in the highest likelihood are chosen. In a model with random effects, however, it is not clear how to deal with differing results for the alternating algorithm. The penalized likelihood can not be used for deciding between two sets of parameter estimates because it depends on the variance parameters. Thus, results with larger variance parameters tend to be preferred. Instead of using a grid of initial values to deal with problems due to poor initial values, convergence may be improved by using a trust-region approach (Dennis and Schnabel, 1996, Chapter 6.4), as implemented for example in the `R` function `nlminb` (Gay, 1983).

The function `nlminb` is now used as default optimizer to update parameters in steps 1 and 2, compare Appendix III. In all applications considered, the algorithm always lead to the same solution after convergence when using different initial values, see Paper III for further details. In the following, the score vectors and Fisher information matrices required in steps 1 and 2 are given.

2.1 Estimation of $\boldsymbol{\beta}$ and $\boldsymbol{b}$

Given known variance components $\boldsymbol{\theta}$ in $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, inference for the parameters $\boldsymbol{\beta}$ and $\boldsymbol{b}$ is based on the penalized log-likelihood

$$\ell_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b};\, \boldsymbol{\Sigma}(\boldsymbol{\theta})) = \ell(\boldsymbol{\beta}, \boldsymbol{b}) + \log p(\boldsymbol{b}|\boldsymbol{\Sigma}(\boldsymbol{\theta}))\,. \tag{12}$$

The corresponding log-likelihood in (12) is, up to additive constants, given as

$$\ell(\boldsymbol{\beta}, \boldsymbol{b}) = \sum_{r,t} l_{rt}$$

with either Poisson log-likelihood contributions

$$l_{rt} = \{y_{rt} \log(\mu_{rt}) - \mu_{rt}\}\ ,$$

or negative binomial log-likelihood contributions

$$\begin{aligned} l_{rt} = \Bigg\{ & \log\Gamma\Big(y_{rt} + \exp(\tilde{\psi})\Big) - \log\Gamma\Big(\exp(\tilde{\psi})\Big) \\ & + \exp(\tilde{\psi}) \log\left(\frac{\exp(\tilde{\psi})}{\exp(\tilde{\psi}) + \mu_{rt}}\right) + y_{rt} \log\left(\frac{\mu_{rt}}{\exp(\tilde{\psi}) + \mu_{rt}}\right) \Bigg\}, \end{aligned}$$

where $\Gamma(\cdot)$ denotes the gamma function (Abramowitz and Stegun, 1964, p. 255).

The vector of random effects $\boldsymbol{b}$ is assumed to follow a multivariate Gaussian distribution with density

$$p(\boldsymbol{b}|\boldsymbol{\Sigma}(\boldsymbol{\theta})) = (2\pi)^{-\dim(\boldsymbol{b})/2} |\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{b}^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{b}\right\}\ ,$$

where $|\cdot|$ denotes the determinant of a matrix. After dropping terms that are constant with respect to $\boldsymbol{b}$, the penalty term $\log p(\boldsymbol{b}|\boldsymbol{\Sigma}(\boldsymbol{\theta}))$ in (12) equals

$$\log(p(\boldsymbol{b}|\boldsymbol{\Sigma}(\boldsymbol{\theta}))) = \boldsymbol{b}^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{b}\,.$$

The score vector with first derivatives of (12) with respect to the fixed and random parameters can be partitioned as

$$\boldsymbol{s}_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{\Sigma}(\boldsymbol{\theta})) = \begin{pmatrix} \frac{\partial \ell_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{\Sigma}(\boldsymbol{\theta}))}{\partial \boldsymbol{\beta}} \\ \frac{\partial \ell_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{\Sigma}(\boldsymbol{\theta}))}{\partial \boldsymbol{b}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{s}(\boldsymbol{\beta}) \\ \boldsymbol{s}(\boldsymbol{b}) - \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{b} \end{pmatrix}\ , \tag{13}$$

where $\boldsymbol{s}(\boldsymbol{i}) = \partial\ell(\boldsymbol{\beta}, \boldsymbol{b})/\partial \boldsymbol{i}$ corresponds to the unpenalized score vector with respect to the parameter vector $\boldsymbol{i}$. Analogously, the observed Fisher information matrix can be partitioned as

$$\begin{aligned} \mathbf{F}_{\text{pen}} = \mathbf{F}_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{\Sigma}(\boldsymbol{\theta})) &= \begin{pmatrix} -\frac{\partial^2 \ell_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{\Sigma}(\boldsymbol{\theta}))}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top} & -\frac{\partial^2 \ell_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{\Sigma}(\boldsymbol{\theta}))}{\partial \boldsymbol{\beta}\partial \boldsymbol{b}^\top} \\ -\frac{\partial^2 \ell_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{\Sigma}(\boldsymbol{\theta}))}{\partial \boldsymbol{b}\partial \boldsymbol{\beta}^\top} & -\frac{\partial^2 \ell_{\text{pen}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{\Sigma}(\boldsymbol{\theta}))}{\partial \boldsymbol{b}\partial \boldsymbol{b}^\top} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{F}[\boldsymbol{\beta}\boldsymbol{\beta}] & \mathbf{F}[\boldsymbol{\beta}\boldsymbol{b}] \\ \mathbf{F}[\boldsymbol{\beta}\boldsymbol{b}] & \mathbf{F}[\boldsymbol{b}\boldsymbol{b}] + \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \end{pmatrix}, \end{aligned} \tag{14}$$

where $\mathbf{F}[\boldsymbol{ij}] = \partial^2 \ell(\boldsymbol{\beta}, \boldsymbol{b}) / (\partial \boldsymbol{i} \partial \boldsymbol{j}^\top)$ denotes the block of the unpenalized Fisher information matrix corresponding to the parameter vectors $\boldsymbol{i}$ and $\boldsymbol{j}$.

Poisson model

For the Poisson distribution, the unpenalized score vector in (13) is given by

$$\boldsymbol{s}(\boldsymbol{\xi}) = \sum_{r,t} \frac{y_{rt}}{\mu_{rt}} \frac{\partial \mu_{rt}}{\partial \boldsymbol{\xi}} - \frac{\partial \mu_{rt}}{\partial \boldsymbol{\xi}} \,,$$

where $\boldsymbol{\xi}$ denotes the stacked vector of all fixed and random effects. The unpenalized Fisher information matrix for the regression parameters in (14) has elements

$$\mathbf{F}[ij] = -\sum_{r,t} \left\{ -\frac{y_{rt}}{\mu_{rt}^2} \frac{\partial \mu_{rt}}{\partial i} \frac{\partial \mu_{rt}}{\partial j} + \frac{y_{rt}}{\mu_{rt}} \frac{\partial^2 \mu_{rt}}{\partial i \partial j} - \frac{\partial^2 \mu_{rt}}{\partial i \partial j} \right\} \,,$$

where i and j denote the i-th and j-th element of $\boldsymbol{\xi}$, respectively.

First and second partial derivatives of μ_{rt} with respect to fixed and random parameters are given by

$$\frac{\partial \mu_{rt}}{\partial \boldsymbol{\xi}^{(\lambda)}} = \lambda_{rt} y_{r,t-1} \, \boldsymbol{u}_{rt}^{(\lambda)} \,, \qquad \frac{\partial \mu_{rt}}{\partial \boldsymbol{\xi}^{(\phi)}} = \phi_{rt} \sum_{q \neq r} w_{qr} y_{q,t-1} \, \boldsymbol{u}_{rt}^{(\phi)} \,, \qquad \frac{\partial \mu_{rt}}{\partial \boldsymbol{\xi}^{(\nu)}} = \nu_{rt} \, \boldsymbol{u}_{rt}^{(\nu)} \,,$$

and

$$\frac{\partial^2 \mu_{rt}}{\partial \boldsymbol{\xi}^{(\lambda)} \partial \boldsymbol{\xi}^{(\lambda)\top}} = \lambda_{rt} y_{r,t-1} \, \boldsymbol{u}_{rt}^{(\lambda)} \boldsymbol{u}_{rt}^{(\lambda)\top} \,, \qquad \frac{\partial^2 \mu_{rt}}{\partial \boldsymbol{\xi}^{(\phi)} \partial \boldsymbol{\xi}^{(\phi)\top}} = \phi_{rt} \sum_{q \neq r} w_{qr} y_{q,t-1} \, \boldsymbol{u}_{rt}^{(\phi)} \boldsymbol{u}_{rt}^{(\phi)\top} \,,$$

$$\frac{\partial^2 \mu_{rt}}{\partial \boldsymbol{\xi}^{(\nu)} \partial \boldsymbol{\xi}^{(\nu)\top}} = \nu_{rt} \, \boldsymbol{u}_{rt}^{(\nu)} \boldsymbol{u}_{rt}^{(\nu)\top} \,.$$

Second partial derivatives of μ_{rt} with respect to parameters belonging to different components are zero.

Negative binomial model

For the negative binomial distribution, the unpenalized score vector in (13) has elements

$$\boldsymbol{s}(\boldsymbol{\xi}) = \sum_{r,t} \left\{ -\frac{\exp(\tilde{\psi})}{\exp(\tilde{\psi}) + \mu_{rt}} + \frac{y_{rt}}{\mu_{rt}} - \frac{y_{rt}}{\exp(\tilde{\psi}) + \mu_{rt}} \right\} \frac{\partial \mu_{rt}}{\partial \boldsymbol{\xi}},$$

$$\boldsymbol{s}(\tilde{\psi}) = \sum_{r,t} \left\{ \Psi\big(y_{rt} + \exp(\tilde{\psi})\big) - \Psi\big(\exp(\tilde{\psi})\big) + \tilde{\psi} + 1 - \log\big(\exp(\tilde{\psi}) + \mu_{rt}\big) - \frac{\exp(\tilde{\psi}) + y_{rt}}{\exp(\tilde{\psi}) + \mu_{rt}} \right\} \exp(\tilde{\psi}),$$

where $\Psi(z) = \frac{d}{dz} \log(\Gamma(z))$ is the digamma function (Abramowitz and Stegun, 1964, p. 258).

The unpenalized Fisher information matrix in (14) has elements

$$\mathbf{F}[ij] = -\sum_{r,t} \left\{ \left(\frac{\exp(\tilde{\psi})}{(\exp(\tilde{\psi}) + \mu_{rt})^2} - \frac{y_{rt}}{\mu_{rt}^2} + \frac{y_{rt}}{(\exp(\tilde{\psi}) + \mu_{rt})^2} \right) \frac{\partial \mu_{rt}}{\partial i} \frac{\partial \mu_{rt}}{\partial j} \right.$$

$$+\left(-\frac{\exp(\tilde{\psi})}{\exp(\tilde{\psi})+\mu_{rt}}+\frac{y_{rt}}{\mu_{rt}}-\frac{y_{rt}}{\exp(\tilde{\psi})+\mu_{rt}}\right)\frac{\partial^2\mu_{rt}}{\partial i\partial j}\Bigg\},$$

$$\mathbf{F}[i\tilde{\psi}]=\mathbf{F}[\tilde{\psi}i]=-\sum_{r,t}\Bigg\{\exp(\tilde{\psi})\left(\frac{-1}{\exp(\tilde{\psi})+\mu_{rt}}+\frac{\exp(\tilde{\psi})+y_{rt}}{(\exp(\tilde{\psi})+\mu_{rt})^2}\right)\frac{\partial\mu_{rt}}{\partial i}\Bigg\},$$

$$\mathbf{F}[\tilde{\psi}\tilde{\psi}]=-\sum_{r,t}\Bigg\{\exp(\tilde{\psi})\Bigg(\Psi'\big(y_{rt}+\exp(\tilde{\psi})\big)\exp(\tilde{\psi})-\Psi'\big(\exp(\tilde{\psi})\big)\exp(\tilde{\psi})+1$$
$$-\frac{\exp(\tilde{\psi})}{\exp(\tilde{\psi})+\mu_{rt}}-\exp(\tilde{\psi})\frac{\mu_{rt}-y_{rt}}{(\exp(\tilde{\psi})+\mu_{rt})^2}\Bigg)\Bigg\}-\boldsymbol{s}(\tilde{\psi})\exp(\tilde{\psi}),$$

where $\Psi'(z)=\frac{d}{dz}\Psi(z)$ is the trigamma function (Abramowitz and Stegun, 1964, p. 260).

2.2 Estimation of variance components $\boldsymbol{\theta}$

Estimates of the variance components in $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ are obtained by maximizing the marginal likelihood

$$L_{\text{marg}}(\boldsymbol{\Sigma}(\boldsymbol{\theta}))=\int\exp\left\{\ell_{\text{pen}}(\boldsymbol{\beta},\boldsymbol{b};\,\boldsymbol{\Sigma}(\boldsymbol{\theta}))\right\}\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\boldsymbol{b}$$
$$\propto|\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{-\frac{1}{2}}\int\exp\left\{\ell(\boldsymbol{\beta},\boldsymbol{b})-\frac{1}{2}\boldsymbol{b}^T\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{b}\right\}\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\boldsymbol{b}$$

with respect to $\boldsymbol{\theta}$. This integral cannot be solved analytically. In Paper III, an approximate marginal likelihood is derived by applying a first-order Laplace approximation to the integrand and assuming that small changes in the variance parameters hardly affect the estimates of the regression coefficients. The resulting log marginal likelihood is given by

$$\ell_{\text{marg}}(\boldsymbol{\Sigma}(\boldsymbol{\theta}))\approx-\frac{1}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})|-\frac{1}{2}\boldsymbol{b}^\top\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{b}-\frac{1}{2}\log|\mathbf{F}_{\text{pen}}(\boldsymbol{\beta},\boldsymbol{b};\boldsymbol{\Sigma}(\boldsymbol{\theta}))|\,, \tag{15}$$

where $\boldsymbol{b}$ and $\boldsymbol{\beta}$ denote fixed values not depending directly on the variance components, e.g. current estimates.

In the following, the random effects vector $\boldsymbol{b}$ is assumed to follow a multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ given in (7) and (8). The use of a block-diagonal covariance matrix (6) results in simplified expressions, compare Appendix I. Furthermore, let k and l denote any element of the vector with variance parameters $\boldsymbol{\theta}=(s_1,s_2,s_3,r_1,r_2,r_3)^\top$.

First note that the log determinant of the covariance matrix is given by

$$\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})|=\log|\boldsymbol{\Omega}(\boldsymbol{\theta})\otimes\mathbf{I}_R|=\log\left(|\mathbf{S}\mathbf{T}\mathbf{T}^\top\mathbf{S}|^R|\,\mathbf{I}_R|^3\right)=\log\left(\{|\mathbf{S}||\mathbf{T}|\}^{2R}\right)$$
$$=2R\left(\sum_{i=1}^3 s_i-\frac{1}{2}\sum_{i=1}^3\log(r_i^2+1)\right).$$

First and second derivatives of (15) with respect to $\boldsymbol{\theta}$ can be derived based on differentiation rules for matrices (Harville, 1997, Chapter 15). For notational convenience, the dependence of the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and the penalized Fisher information matrix $\mathbf{F}_{\text{pen}}(\boldsymbol{\beta},\boldsymbol{b};\boldsymbol{\Sigma}(\boldsymbol{\theta}))$ on regression and variance parameters will be suppressed in the following. The score vector

$\boldsymbol{s}_{\text{marg}}(\boldsymbol{\Sigma})$ has elements

$$\frac{\partial \ell_{\text{marg}}(\boldsymbol{\Sigma})}{\partial k} = -R\sum_{i=1}^{3}\left(\mathbb{1}(k=s_i) - \frac{r_i}{r_i^2+1}\mathbb{1}(k=r_i)\right) + \frac{1}{2}\boldsymbol{b}^\top\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial k}\boldsymbol{\Sigma}^{-1}\boldsymbol{b} - \frac{1}{2}\operatorname{tr}\left(\mathbf{F}_{\text{pen}}^{-1}\frac{\partial\mathbf{F}_{\text{pen}}}{\partial k}\right)$$

where $\operatorname{tr}(\cdot)$ denotes the trace of a matrix. The elements of the observed Fisher information matrix $\mathbf{F}_{\text{marg}}(\boldsymbol{\Sigma})$ are given as

$$\begin{aligned}\frac{\partial^2 \ell_{\text{marg}}(\boldsymbol{\Sigma})}{\partial k \partial l} =& -R\sum_{i=1}^{3}\left(\frac{r_i^2-1}{(r_i^2+1)^2}\mathbb{1}(k=l=r_i)\right)\\ &-\frac{1}{2}\boldsymbol{b}^\top\boldsymbol{\Sigma}^{-1}\left(-\frac{\partial^2\boldsymbol{\Sigma}}{\partial k\partial l} + \frac{\partial\boldsymbol{\Sigma}}{\partial k}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial l} + \frac{\partial\boldsymbol{\Sigma}}{\partial l}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial k}\right)\boldsymbol{\Sigma}^{-1}\boldsymbol{b}\\ &-\frac{1}{2}\operatorname{tr}\left(-\mathbf{F}_{\text{pen}}^{-1}\frac{\partial\mathbf{F}_{\text{pen}}}{\partial l}\mathbf{F}_{\text{pen}}^{-1}\frac{\partial\mathbf{F}_{\text{pen}}}{\partial k} + \mathbf{F}_{\text{pen}}^{-1}\frac{\partial^2\mathbf{F}_{\text{pen}}}{\partial k\partial l}\right).\end{aligned}$$

First and second derivatives of $\boldsymbol{\Sigma}$ and $\mathbf{F}_{\text{pen}}$

First recall the partition of the penalized Fisher information matrix (14) into blocks with respect to the fixed effects $\boldsymbol{\beta}$ and the random effects $\boldsymbol{b}$. Only the block of $\mathbf{F}_{\text{pen}}$ corresponding to $\boldsymbol{b}$ involves variance parameters $\boldsymbol{\theta}$. The non-zero first and second derivatives of the penalized Fisher information matrix $\mathbf{F}_{\text{pen}}$ with respect to $\boldsymbol{\theta}$ are thus given by

$$\frac{\partial\mathbf{F}_{\text{pen}}[\boldsymbol{bb}]}{\partial k} = \frac{\partial\mathbf{F}[\boldsymbol{bb}]}{\partial k} + \frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial k} = -\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial k}\boldsymbol{\Sigma}^{-1}$$

and

$$\frac{\partial^2\mathbf{F}_{\text{pen}}[\boldsymbol{bb}]}{\partial k\partial l} = \boldsymbol{\Sigma}^{-1}\left(-\frac{\partial^2\boldsymbol{\Sigma}}{\partial k\partial l} + \frac{\partial\boldsymbol{\Sigma}}{\partial k}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial l} + \frac{\partial\boldsymbol{\Sigma}}{\partial l}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial k}\right)\boldsymbol{\Sigma}^{-1}.$$

First and second derivatives of the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Omega}(\boldsymbol{\theta}) \otimes \mathbf{I}_R$ are given by

$$\frac{\partial\boldsymbol{\Sigma}}{\partial k} = \frac{\partial\boldsymbol{\Omega}(\boldsymbol{\theta})}{\partial k}\otimes\mathbf{I}_R \qquad \text{and} \qquad \frac{\partial^2\boldsymbol{\Sigma}}{\partial k\partial l} = \frac{\partial^2\boldsymbol{\Omega}(\boldsymbol{\theta})}{\partial k\partial l}\otimes\mathbf{I}_R.$$

Let Ω_{ij} denote the (i,j)-th element of the matrix $\boldsymbol{\Omega}(\boldsymbol{\theta})$ and let $\partial_l\Omega_{ij}$ and $\partial_{kl}\Omega_{ij}$ denote the first and second partial derivatives of Ω_{ij} with respect to k, and k,l, respectively. Furthermore, let $\tilde{r}_i := \sqrt{r_i^2+1}$, $i=1,2,3$.

The non-zero first partial derivatives of the upper triangular elements of the symmetric matrix $\boldsymbol{\Omega}(\boldsymbol{\theta})$ are given by:

$$\begin{aligned}&\partial_{s_i}\Omega_{ii} = 2\exp(2s_i), \quad i=1,2,3 && \partial_{s_i}\Omega_{12} = \frac{r_1}{\tilde{r}_1}\exp(s_1+s_2), \quad i=1,2\\ &\partial_{s_i}\Omega_{13} = \frac{r_2}{\tilde{r}_2}\exp(s_1+s_3), \quad i=1,3 && \partial_{s_i}\Omega_{23} = \frac{r_1r_2\tilde{r}_3+r_3}{\tilde{r}_1\tilde{r}_2\tilde{r}_3}\exp(s_2+s_3) \quad i=2,3\\ &\partial_{r_1}\Omega_{12} = (\tilde{r}_1)^{-3}\exp(s_1+s_2) && \partial_{r_1}\Omega_{23} = \frac{r_2\tilde{r}_3 - r_1r_3}{(\tilde{r}_1)^3\tilde{r}_2\tilde{r}_3}\exp(s_2+s_3)\end{aligned}$$

$$\partial_{r_2}\Omega_{13} = (\tilde{r}_2)^{-3}\exp(s_1+s_3) \qquad \partial_{r_2}\Omega_{23} = \frac{r_1\tilde{r}_3 - r_2 r_3}{\tilde{r}_1(\tilde{r}_2)^3\tilde{r}_3}\exp(s_2+s_3)$$

$$\partial_{r_3}\Omega_{23} = \frac{1}{\tilde{r}_1\tilde{r}_3(\tilde{r}_2)^3}\exp(s_2+s_3)$$

The non-zero second partial derivatives of the upper triangular elements of the symmetric matrix $\boldsymbol{\Omega}(\boldsymbol{\theta})$ are given by:

$$\partial_{s_i s_i}\Omega_{ii} = 4\exp(2s_i)\,, \quad i=1,2,3$$

$$\partial_{s_1 s_i}\Omega_{12} = \frac{r_1}{\tilde{r}_1}\exp(s_1+s_2)\,, \quad i=1,2 \qquad \partial_{s_1 s_i}\Omega_{13} = \frac{r_2}{\tilde{r}_2}\exp(s_1+s_3)\,, \quad i=1,3$$

$$\partial_{s_1 r_1}\Omega_{12} = (\tilde{r}_1)^{-3}\exp(s_1+s_2) \qquad \partial_{s_1 r_2}\Omega_{13} = (\tilde{r}_2)^{-3}\exp(s_1+s_3)$$

$$\partial_{s_2 s_2}\Omega_{12} = \frac{r_1}{\tilde{r}_1}\exp(s_1+s_2) \qquad \partial_{s_2 s_i}\Omega_{23} = \frac{r_1 r_2\tilde{r}_3 + r_3}{\tilde{r}_1\tilde{r}_2\tilde{r}_3}\exp(s_2+s_3)\,, \quad i=2,3$$

$$\partial_{s_2 r_1}\Omega_{12} = (\tilde{r}_1)^{-3}\exp(s_1+s_2) \qquad \partial_{s_2 r_1}\Omega_{23} = \frac{r_2\tilde{r}_3 - r_1 r_3}{(\tilde{r}_1)^3\tilde{r}_2\tilde{r}_3}\exp(s_2+s_3)$$

$$\partial_{s_2 r_2}\Omega_{23} = \frac{r_1\tilde{r}_3 - r_2 r_3}{\tilde{r}_1(\tilde{r}_2)^3\tilde{r}_3}\exp(s_2+s_3) \qquad \partial_{s_2 r_3}\Omega_{23} = \frac{1}{\tilde{r}_1\tilde{r}_2(\tilde{r}_3)^3}\exp(s_2+s_3)$$

$$\partial_{s_3 s_3}\Omega_{13} = \frac{r_2}{\tilde{r}_2}\exp(s_1+s_3) \qquad \partial_{s_3 s_3}\Omega_{23} = \frac{r_1 r_2\tilde{r}_3 + r_3}{\tilde{r}_1\tilde{r}_2\tilde{r}_3}\exp(s_2+s_3)$$

$$\partial_{s_3 r_1}\Omega_{23} = \frac{r_2\tilde{r}_3 - r_1 r_3}{(\tilde{r}_1)^3\tilde{r}_2\tilde{r}_3}\exp(s_2+s_3) \qquad \partial_{s_3 r_2}\Omega_{13} = (\tilde{r}_2)^{-3}\exp(s_1+s_3)$$

$$\partial_{s_3 r_2}\Omega_{23} = \frac{r_1\tilde{r}_3 - r_2 r_3}{\tilde{r}_1(\tilde{r}_2)^3\tilde{r}_3}\exp(s_2+s_3) \qquad \partial_{s_3 r_3}\Omega_{23} = \frac{1}{\tilde{r}_1\tilde{r}_2(\tilde{r}_3)^3}\exp(s_2+s_3)$$

$$\partial_{r_1 r_1}\Omega_{12} = -\frac{3r_1}{(\tilde{r}_1)^5}\exp(s_1+s_2) \qquad \partial_{r_1 r_1}\Omega_{23} = -\frac{3r_1 r_2\tilde{r}_3 - 2r_1^2 r_3 + r_3}{(\tilde{r}_1)^5\tilde{r}_2\tilde{r}_3}\exp(s_2+s_3)$$

$$\partial_{r_1 r_2}\Omega_{23} = \frac{\tilde{r}_3 + r_1 r_2 r_3}{(\tilde{r}_1)^3(\tilde{r}_2)^3\tilde{r}_3}\exp(s_2+s_3) \qquad \partial_{r_1 r_3}\Omega_{23} = -\frac{r_1}{(\tilde{r}_1)^3\tilde{r}_2(\tilde{r}_3)^3}\exp(s_2+s_3)$$

$$\partial_{r_2 r_2}\Omega_{13} = -\frac{3r_2}{(\tilde{r}_2)^5}\exp(s_1+s_3) \qquad \partial_{r_2 r_2}\Omega_{23} = -\frac{3r_1 r_2\tilde{r}_3 - 2r_2^2 r_3 + r_3}{\tilde{r}_1(\tilde{r}_2)^5\tilde{r}_3}\exp(s_2+s_3)$$

$$\partial_{r_2 r_3}\Omega_{23} = -\frac{r_2}{\tilde{r}_1(\tilde{r}_2)^3(\tilde{r}_3)^3}\exp(s_2+s_3) \qquad \partial_{r_3 r_3}\Omega_{23} = -\frac{3r_3}{\tilde{r}_1\tilde{r}_2(\tilde{r}_3)^5}\exp(s_2+s_3)$$

References

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York.

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics, *Annals of the Institute of Statistical Mathematics* **43**(1): 1–20.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**(421): 9–25.

Dennis, J. E. and Schnabel, R. B. (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Society for Industrial and Applied Mathematics, Philadelphia.

Gay, D. M. (1983). Algorithm 611. Subroutines for unconstrained minimization using a model/trust-region approach, *ACM Transactions on Mathematical Software* **9**(4): 503–524.

Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*, Springer, New York.

Kneib, T. and Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression, *Scandinavian Journal of Statistics* **34**(1): 207–228.

Paul, M. and Held, L. (2009). Incorporating random effects in a likelihood-based model for multivariate infectious disease counts, *Proceedings of the* 16^{th} *European Young Statisticians meeting*, Bucharest, Romania.

Paul, M. and Held, L. (2010). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Statistics in Medicine*. Accepted.

Paul, M., Held, L. and Toschke, A. M. (2008). Multivariate modelling of infectious disease surveillance data, *Statistics in Medicine* **27**(29): 6250–6267.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model, *Journal of Computational and Graphical Statistics* **4**(1): 12–35.

Pinheiro, J. C. and Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices, *Statistics and Computing* **6**(3): 289–296.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. `http://www.R-project.org`.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature, *The STATA Journal* **2**: 1–21.

Rapisarda, F., Brigo, D. and Mercurio, F. (2007). Parameterizing correlations: a geometric interpretation, *IMA Journal of Management Mathematics* **18**(1): 55–73.

Raudenbush, S. W., Yang, M. L. and Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation, *Journal of Computational and Graphical Statistics* **9**(1): 141–157.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*, Vol. 104, Chapman & Hall/CRC Press, London.

Serfling, R. (1963). Methods for current statistical analysis of excess pneumonia-influenza deaths, *Public Health Reports* **78**(6): 494–506.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association* **81**(393): 82–86.

Wolfinger, R. D. (1999). Fitting nonlinear mixed models with the new NLMIXED procedure (Paper 287), *Proceedings of the 24th SAS User's Group Meeting.*, Cary, NC: SAS Institute.

APPENDIX III

Software manual

The function 'hhh4' in the R-package 'surveillance'

Michaela Paul*
Biostatistics Unit
Institute of Social and Preventive Medicine
University of Zurich, Switzerland

Abstract

This document gives an introduction to the use of the function `hhh4` for modelling univariate and multivariate time series of infectious disease counts. The function is part of the R-package `surveillance`, which provides tools for the visualization, modelling and monitoring of surveillance time series. The basic functionality of `surveillance` is introduced in the package vignette (Höhle et al., 2007) and in Höhle (2007) with main focus on outbreak detection methods. The following illustrates the use of `hhh4` as estimation and prediction routine for the modelling framework proposed by Held et al. (2005), and extended in Paul et al. (2008), Paul and Held (2010) and Herzog et al. (2010).

1 Introduction

To meet the threats of infectious diseases, many countries have established surveillance systems for the reporting of various infectious diseases. The systematic and standardized reporting at a national and regional level aims to recognize all outbreaks quickly, even when aberrant cases are dispersed in space. Traditionally, notification data, i.e. counts of cases confirmed according to a specific definition and reported daily, weekly or monthly on a regional or national level, are used for surveillance purposes.

The R-package `surveillance` provides functionality for the retrospective modelling and prospective change-point detection in the resulting surveillance time series. A recent introduction to the package with focus on outbreak detection methods is given by Höhle and Mazick (2010).

This document illustrates the functionality of the function `hhh4` for the modelling of univariate and multivariate time series of infectious disease counts. The function is currently incorporated in the development version of `surveillance` available from `http://surveillance.r-forge.r-project.org/`. Section 2 introduces the S4 class data structure used to store surveillance time series data within the package. Access and visualization methods

*Author of correspondence: Email: `michaela.paul@ifspm.uzh.ch`

are outlined by means of built-in data sets. In Section 3, the statistical modelling approach by Held et al. (2005) and further model extensions are described. After the general function call and arguments are shown, the detailed usage of `hhh4` is demonstrated in Section 4 using data introduced in Section 2.

2 Surveillance data

Denote by $\{y_{it}; i = 1, \ldots, I, t = 1, \ldots, T\}$ the multivariate time series of disease counts for a specific partition of gender, age and location. Here, T denotes the length of the time series and I denotes the number of units (e.g geographical regions or age groups) being monitored. Such data are represented using objects of the S4 class `sts` (surveillance time series).
This class contains the $T \times I$ matrix of counts y_{it} in a slot `observed`. An integer slot `epoch` denotes the time index $1 \leq t \leq T$ of each row in `observed`. The number of observations per year, e.g. 52 for weekly or 12 for monthly data, is denoted by `freq`. Furthermore, `start` denotes a vector of length two containing the start of the time series as `c(year, epoch)`. For spatially stratified time series, the slot `neighbourhood` denotes an $I \times I$ adjacency matrix with elements 1 if two regions are neighbors and 0 otherwise. For map visualizations, the slot `map` links the multivariate time series to geographical regions of an ESRI shapefile (using functionality from the package `maptools` (Lewin-Koh et al., 2010)). Additionally, the slot `populationFrac` contains a $T \times I$ matrix representing population fractions in unit i at time t.
The package `surveillance` contains a number of time series in the `data` directory. Most data sets originate from the SurvStat@RKI database (`http://www3.rki.de/SurvStat`), maintained by the Robert Koch Institute (RKI), Germany. Selected data sets will be analyzed in Section 4 and are introduced in the following. Note that many of the built-in datasets are stored in the S3 class data structure `disProg`. They can be easily converted into the S4 `sts` data structure using the function `disProg2sts`. The resulting `sts` object can be accessed similar as standard `matrix` objects and allows easy temporal and spatial aggregation as will be shown in the remainder of this section.

Example: Influenza and meningococcal disease in Germany 01/2001–52/2006
As a first example, the weekly number of influenza and meningococcal disease cases in Germany is considered.

```
> data(influMen)
> # convert to sts class and print basic information about the time series
> print(fluMen <- disProg2sts(influMen))
```

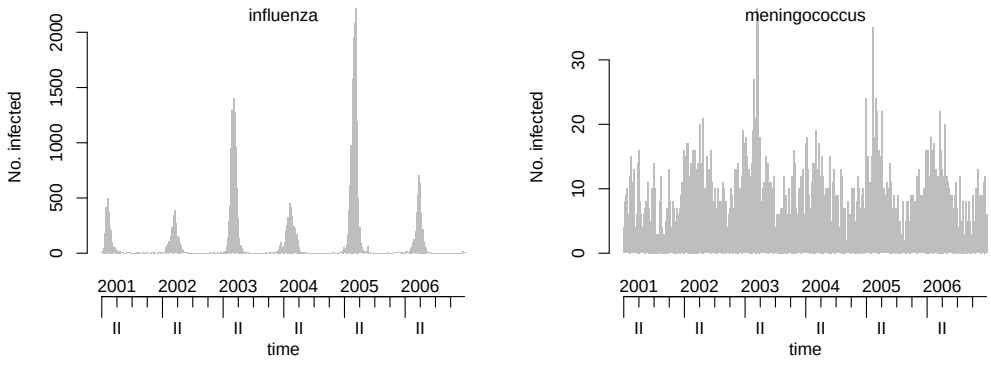
```
-- An object of class sts --
freq:                 52
start:                 2001 1
dim(observed):         312 2

Head of observed:
     influenza meningococcus
[1,]         7             4

map:
NULL

head of neighbourhood:
          influenza meningococcus
influenza         0             1
```

The univariate time series of meningococcal disease counts can be obtained with

```
> meningo <- fluMen[, "meningococcus"]
> dim(meningo)

[1] 312   1
```

The `plot` function provides an interface to the visual representation of the multivariate time series in time, space and space-time which is controlled by the `type` argument:

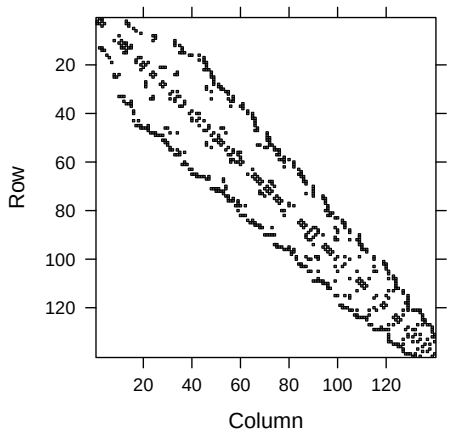
```
> plot(fluMen, type = observed ~ time | unit, # type of plot
+              same.scale = FALSE,            # unit-specific ylim ?
+              col = "grey"                   # color of bars
+              )
```



See Höhle and Mazick (2010) for a detailed description of the plot routines.

Example: Influenza in Southern Germany, 01/2001-52/2008 The spatio-temporal spread of influenza in the 140 Kreise (districts) of Bavaria and Baden-Württemberg is analyzed using the weekly number of cases reported to the RKI (Robert Koch-Institut, 2009) in the years 2001–2008. An `sts` object containing the data is created as follows:

```
> flu.counts <- as.matrix(read.table("../data/flu_ByBw.txt"))
> # read in adjacency matrix with elements 1 if two regions share a common border
> nhood <- as.matrix(read.table("../data/neighourhood_ByBw.txt"))
> # visualize adjacency matrix
> image(Matrix(nhood))
```

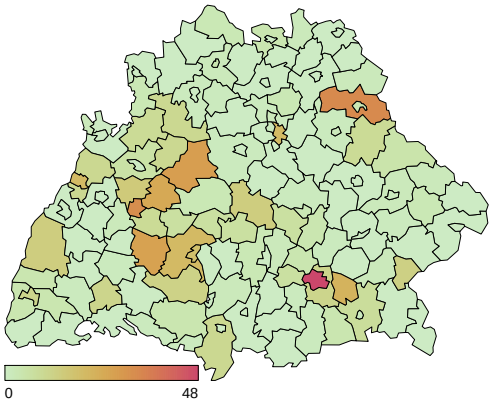


Dimensions: 140 x 140

```
> map <- readShapePoly("../shapes/districts_BYBW.shp", IDvar = "id")
> # read in population fractions
> p <- as.matrix(read.table("../data/population_2001-12-31_ByBw.txt"))
> # create sts object
> flu <- new("sts", epoch = 1:nrow(flu.counts),
+                   observed = flu.counts,
+                   start = c(2001, 1),
+                   freq = 52,
+                   neighbourhood = nhood,
+                   map = map,
+                   population = p
+                   )
```

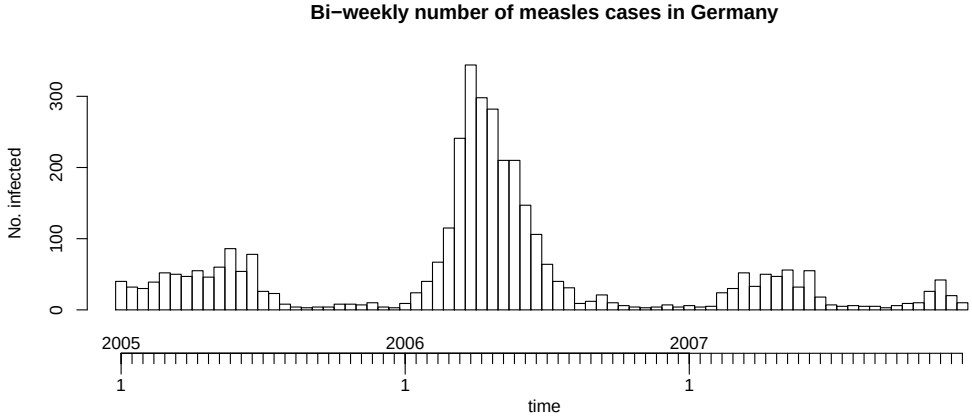
A map of the total number of cases in the year 2001 may be obtained as follows:

```
> plot(flu[year(flu) == 2001, ],    # select year 2001
+      type = observed ~ 1 | unit,  # map of counts aggregated over times t
+      labels = FALSE               # suppress region labels in map
+      )
```



Example: Measles in Germany, 01/2005–52/2007 The following data set contains the weekly number of measles cases in the 16 German Bundesländer (federal states), in the years 2005–2007. These data are analyzed in Herzog et al. (2010) after aggregation into successive bi-weekly periods.

```
> data(measles.land)
> # aggregate into successive bi-weekly periods
> measles2w <- aggregate(measles.land, nfreq = 26)
> plot(measles2w, type = observed ~ time,   # ts plot aggregated over all units i
+                 main = "Bi-weekly number of measles cases in Germany",
+                 legend.opts = NULL        # suppress default legend
+                 )
```



3 Model formulation

Retrospective surveillance aims to identify outbreaks and (spatio-)temporal patterns through statistical modelling. Motivated by a branching process with immigration, Held et al. (2005) suggest the following model for the analysis of univariate time series of infectious disease counts $\{y_t; t = 1, \ldots, T\}$.

The counts are assumed to be Poisson distributed with conditional mean

$$\mu_t = \lambda y_{t-1} + \nu_t, \quad (\lambda, \nu_t > 0)$$

where λ and ν_t are unknown quantities. The mean incidence is decomposed additively into two components: an epidemic or *autoregressive* component λy_{t-1}, and an *endemic* component ν_t. The former should be able to capture occasional outbreaks whereas the latter explains a baseline rate of cases with stable temporal pattern. Held et al. (2005) suggest the following parametric model for the endemic component:

$$\log(\nu_t) = \alpha + \beta t + \left\{ \sum_{s=1}^{S} \gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t) \right\}, \tag{1}$$

where α is an intercept, β is a trend parameter, and the terms in curly brackets are used to model seasonal variation. Here, γ_s and δ_s are unknown parameters, S denotes the number of harmonics to include, and $\omega_s = 2\pi s/\texttt{freq}$ are Fourier frequencies (e.g. `freq = 52` for weekly data). For ease of interpretation, the seasonal terms in (1) can be written equivalently as

$$\gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t) = A_s \sin(\omega_s t + \varphi_s)$$

with amplitude $A_s = \sqrt{\gamma_s^2 + \delta_s^2}$ describing the magnitude, and phase difference $\tan(\varphi_s) = \delta_s/\gamma_s$ describing the onset of the sine wave.
To account for overdispersion, the Poisson model may be replaced by a negative binomial model. Then, the conditional mean μ_t remains the same but the conditional variance increases to $\mu_t(1 + \mu_t/\psi)$ with additional unknown overdispersion parameter $\psi > 0$.
The model is extended to multivariate time series $\{y_{it}\}$ in Held et al. (2005) and Paul et al. (2008) by including an additional *neighbor-driven* component, where past cases in other (neighboring) units also enter as explanatory covariates. The conditional mean μ_{it} is then given by

$$\mu_{it} = \lambda y_{i,t-1} + \phi \sum_{j \neq i} w_{ji} y_{j,t-1} + e_{it}\nu_t, \tag{2}$$

where the unknown parameter ϕ quantifies the influence of other units j on unit i, w_{ji} are suitably chosen known weights and e_{it} corresponds to an offset (such as population fractions at time t in region i). A simple choice for the weights is $w_{ji} = 1$ if units j and i are adjacent and 0 otherwise. See Paul et al. (2008) for a discussion of alternative weights.
When analyzing a specific disease observed in, say, multiple regions or several pathogens (such as influenza and meningococcal disease), the assumption of equal incidence levels or disease transmission across units is questionable. To address such heterogeneity, the unknown quantities λ, ϕ, and ν_t in (2) may also depend on unit i. This can be done via

- unit-specific fixed parameters, e.g. $\log(\lambda_i) = \alpha_i$ (Paul et al., 2008);
- unit-specific random effects, e.g $\log(\lambda_i) = \alpha_0 + a_i$, $a_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_\lambda)$ (Paul and Held, 2010);
- linking parameters with known (possibly time-varying) explanatory variables, e.g. $\log(\lambda_i) = \alpha_0 + x_i\alpha_1$ with region-specific vaccination coverage x_i (Herzog et al., 2010).

A call to `hhh4` fits a Poisson or negative binomial model with conditional mean

$$\mu_{it} = \lambda_{it} y_{i,t-1} + \phi_{it} \sum_{j \neq i} w_{ji} y_{j,t-1} + e_{it}\nu_{it}$$

to a multivariate time series of counts. Here, the three unknown quantities are decomposed additively on the log scale

$$\log(\lambda_{it}) = \alpha_0 + a_i + \boldsymbol{u}_{it}^\top \boldsymbol{\alpha} \qquad (\texttt{ar})$$
$$\log(\phi_{it}) = \beta_0 + b_i + \boldsymbol{x}_{it}^\top \boldsymbol{\beta} \qquad (\texttt{ne})$$
$$\log(\nu_{it}) = \gamma_0 + c_i + \boldsymbol{z}_{it}^\top \boldsymbol{\gamma} \qquad (\texttt{end})$$

where $\alpha_0, \beta_0, \gamma_0$ are intercepts, $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ are vectors of unknown parameters corresponding to covariate vectors $\boldsymbol{u}_{it}, \boldsymbol{x}_{it}, \boldsymbol{z}_{it}$, and a_i, b_i, c_i are random effects. For instance, model (1) with $S = 1$ seasonal terms may be represented as $\boldsymbol{z}_{it} = (t, \sin(2\pi/\texttt{freq}\, t), \cos(2\pi/\texttt{freq}\, t))^\top$. The stacked vector of all random effects is assumed to follow a normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, see Paul and Held (2010) for further details. Inference is based on (penalized) likelihood methodology as proposed in Paul and Held (2010). In applications, each component (`ar`)–(`end`) may be omitted in parts or as a whole.

4 Function call and control settings

The estimation procedure is called with

```
> hhh4(sts, control)
```

where `sts` denotes a (multivariate) surveillance time series and the model is specified in the argument `control` in consistency with other algorithms in `surveillance`. The `control` setting is a list of the following arguments:

```
> control = list(
+     ar = list(f = ~ -1),       # formula: exp(u'alpha) * y_i,t-1
+     ne = list(f = ~ -1,        # formula: exp(x'beta) * sum_j {w_ji * y_j,t-1}
+               weights = NULL   # matrix with weights w_ji
+                                # [w_ji = neighbourhood(stsObj) as default]
```

```
+              ),
+     end = list(f = ~ 1,         # formula:  exp(z'gamma) * e_it
+                offset = NULL    # optional offset e_it
+                ),
+     family = "Poisson",                # Poisson or NegBin model
+     subset = 2:nrow(stsObj),           # subset of observations to be used
+                                        # in the fitting process
+     optimizer = list(tech = "nlminb"), # details for optimizer
+     verbose = FALSE,                   # no progress information is printed
+     start = list(fixed = NULL,         # list with initial values for fixed,
+                  random = NULL,        # random, and
+                  sd.corr = NULL        # variance parameters
+                  ),
+     data = data.frame(t = epoch(sts))  # data.frame,
+                                        # or named list with covariates
+     )
```

The first three arguments `ar`, `ne`, and `end` specify the model components using `formula` objects. As default, the counts y_{it} are assumed to be Poisson distributed. A negative binomial model is obtained with `family = "NegBin1"`. The log-likelihood is maximized using the optimization routine implemented in `nlminb`. Alternatively, the methods implemented in `optim` may be used, e.g. `optimizer = list(tech = "BFGS")`. Initial values for the fixed, random, and variance parameters are passed on in the `start` argument. If the model contains covariates, these have to be specified in the `data` argument. When covariates do not vary across units, they may be passed on as a vector of length T. Otherwise, covariate values have to be stored and passed on in a matrix of size $T \times I$.
In the following, the functionality of `hhh4` is demonstrated using the data sets introduced in Section 2 and previously analyzed in Paul et al. (2008), Paul and Held (2010) and Herzog et al. (2010). Selected results are reproduced. For a thorough discussion we refer to these papers.

Univariate modelling

As a first example, consider the univariate time series of meningococcal infections in Germany, 01/2001–52/2006 (cf. Tab. 1 in Paul et al., 2008). A Poisson model without autoregression and $S = 1$ seasonal term is specified as follows:

```
> ( f_S1 <- addSeason2formula(f = ~ 1, S = 1, period = 52) )


~1 + sin(2 * pi * t/52) + cos(2 * pi * t/52)
<environment: 0xdc21e08>


> # fit Poisson model
> hhh4(meningo, control = list(end = list(f = f_S1), family = "Poisson"))
```

```
Call:
hhh4(stsObj = meningo, control = list(end = list(f = f_S1), family = "Poisson"))


Fixed effects:
                        Estimates  Std.Error
end.1                      2.2648     0.0187
end.sin(2 * pi * t/52)     0.3619     0.0259
end.cos(2 * pi * t/52)     0.2605     0.0258

log-likelihood:    -872.09
AIC:               1750.19
BIC:               1761.41

Number of units:          1
Number of time points:    311
```

A corresponding negative binomial model is obtained via

```
> result1 <- hhh4(meningo, control = list(end = list(f = f_S1),
+                                         family = "NegBin1"))
```

As default, the autoregressive component is omitted with $\sim$ `-1` in the formula specification. In can be included in the model with

```
> m2 <- list(ar = list(f = ~ 1),     # log(lambda) = alpha
+            end = list(f = f_S1),
+            family = "NegBin1",
+            # use estimates from previous model as initial values
+            start = list(fixed = c(log(0.1),      # initial values for alpha,
+                                   coef(result1)) # and remaining parameters
+                         )
+            )
> # fit model
> result2 <- hhh4(meningo, control = m2)
> # extract ML estimates
> round(coef(result2, se = TRUE,      # also return standard errors
+                     idx2Exp = 1     # exponentiate 1st param [-> exp(alpha)]
+                     ),2)


                       Estimates Std. Error
exp(ar.1)                  0.16       0.06
end.1                      2.09       0.07
end.sin(2 * pi * t/52)     0.34       0.04
end.cos(2 * pi * t/52)     0.26       0.04
1/overdisp                 0.05       0.01


> # get AIC
> AIC(result2)


[1] 1701.228
```

Bivariate modelling

Now, the weekly numbers of both meningococcal disease (MEN) and influenza (FLU) cases are analyzed to investigate whether influenza infections predispose meningococcal disease (cf. Tab. 2 in Paul et al., 2008). This requires disease-specific parameters which are specified in the formula object with `fe(...)`. In the following, a negative binomial model with mean

$$\begin{pmatrix}\mu_{\text{men},t}\\ \mu_{\text{flu},t}\end{pmatrix} = \begin{pmatrix}\lambda_{\text{men}} & \phi\\ 0 & \lambda_{\text{flu}}\end{pmatrix}\begin{pmatrix}\text{MEN}_{t-1}\\ \text{FLU}_{t-1}\end{pmatrix} + \begin{pmatrix}\nu_{\text{men},t}\\ \nu_{\text{flu},t}\end{pmatrix},$$

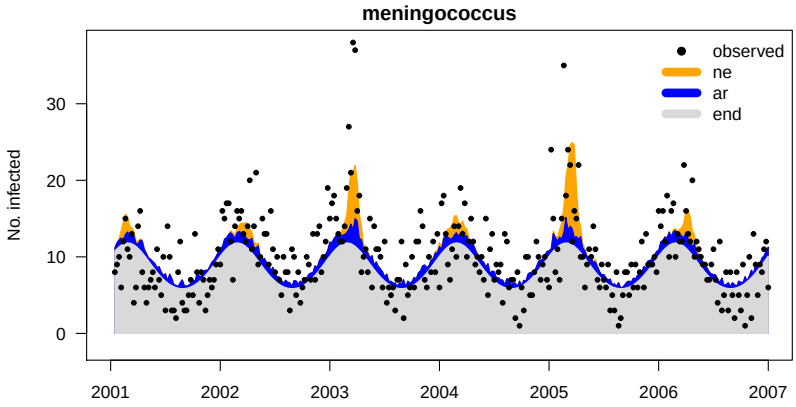
where the endemic component includes $S = 3$ seasonal terms for the FLU data and $S = 1$ seasonal terms for the MEN data is considered. Here, ϕ quantifies the influence of past influenza cases on the meningococcal disease incidence. This model corresponds to the second model of Tab. 2 in Paul et al. (2008) and is fitted with

```
> f.end <- addSeason2formula(f = ~ -1 + fe(1, which = c(TRUE, TRUE)),
+                                          # disease-specific intercepts
+                            S = c(3, 1),    # S = 3 for flu, S = 1 for men
+                            period = 52)
> # specify model
> m <- list(ar = list(f = ~ -1 + fe(1, which=c(TRUE, TRUE))),
+           ne = list(f = ~ -1 + fe(1, which=c(FALSE, TRUE))),
+           end = list(f = f.end),
+           family = "NegBinM"
+           )
> # fit model
> (result <- hhh4(fluMen, control = m))


Call:
hhh4(stsObj = fluMen, control = m)


Fixed effects:
                                       Estimates  Std.Error
ar.1.influenza                           -0.3044     0.0678
ar.1.meningococcus                       -2.3523     0.5980
ne.1.meningococcus                       -5.2167     0.2605
end.1.influenza                           1.0883     0.1653
end.1.meningococcus                       2.1186     0.0668
end.sin(2 * pi * t/52).influenza          1.1862     0.2360
end.sin(2 * pi * t/52).meningococcus      0.2666     0.0397
end.cos(2 * pi * t/52).influenza          1.5098     0.1467
end.cos(2 * pi * t/52).meningococcus      0.2290     0.0353
end.sin(4 * pi * t/52).influenza          0.9192     0.1715
end.cos(4 * pi * t/52).influenza         -0.1616     0.1799
end.sin(6 * pi * t/52).influenza          0.3692     0.1500
end.cos(6 * pi * t/52).influenza         -0.5345     0.1619
1/overdisp.influenza                      0.2946     0.0358
1/overdisp.meningococcus                  0.0395     0.0109

log-likelihood:    -1880.97
AIC:               3791.94
BIC:               3858.43
```

```
Number of units:          2
Number of time points:    311
```

A plot of the estimated mean for the meningococcal disease data, decomposed into the three components, is obtained with

```
> plot(result, i = 2, col = c("orange", "blue", "grey85"), legend = TRUE)
```



Multivariate modelling

For disease counts observed in a large number of regions, say, (i.e. highly multivariate time series of counts) the use of region-specific parameters to account for regional heterogeneity is no longer feasible, as estimation and identifiability problems may occur. Paul and Held (2010) propose a random effects formulation to analyze the weekly number of influenza cases in 140 districts of Southern Germany. For example, consider a model with random intercepts in the endemic component: $c_i \sim \mathcal{N}(0, \sigma_\nu^2), i = 1, \ldots, I$. Such effects are specified in a formula object as

```
> f.end <- ~ -1 + ri(type = "iid", corr = "all")
```

Setting `type = "car"` would assume that the random effects are spatially correlated instead of uncorrelated. See Paul and Held (2010) for further details. The argument `corr = "all"` allows for correlation between region-specific random effects in different components, e.g. random incidence levels c_i in the endemic component and random effects b_i in the neighbor-driven component. The following call to `hhh4` fits such a random effects model with linear trend and $S = 3$ seasonal terms in the endemic component and a fixed autoregressive parameter λ to the influenza data (cf. model B2 in Tab. 3 in Paul and Held, 2010).

```
> wji <- neighbourhood(flu)/rowSums(neighbourhood(flu))
> # endemic component: iid random effects, linear trend, and S=3 seasonal terms
> f.end <- addSeason2formula(f = ~ -1 + ri(type = "iid", corr="all") +
+                                  I((t-208)/100),
+                                  S = 3,
+                                  period = 52)
> model.B2 <- list(ar = list(f = ~ 1),
+                  ne = list(f = ~ -1+ ri(type = "iid", corr="all"),
+                            weights = wji),
+                  end = list(f = f.end, offset = population(flu)),
+                  family = "NegBin1"
+                  )
> # fit model
> (result.B2 <- hhh4(flu, model.B2))


Call:
hhh4(stsObj = flu, control = model.B2)

Random effects:
             Var    Corr
ne.ri(iid)  0.9594
end.ri(iid) 0.5094 0.5617


Fixed effects:
                        Estimates  Std.Error
ar.1                      -0.8976     0.0369
ne.ri(iid)                -1.5256     0.1035
end.I((t - 208)/100)       0.5620     0.0235
end.sin(2 * pi * t/52)     2.1849     0.0985
end.cos(2 * pi * t/52)     2.3319     0.1224
end.sin(4 * pi * t/52)     0.4403     0.1053
end.cos(4 * pi * t/52)    -0.3947     0.0940
end.sin(6 * pi * t/52)     0.3217     0.0648
end.cos(6 * pi * t/52)    -0.2647     0.0631
end.ri(iid)                0.2192     0.1028
1/overdisp                 1.0991     0.0343

penalized log-likelihood:  -18742.42
marginal log-likelihood:   -343.26

Number of units:          140
Number of time points:    416
```

Model choice based on information criteria such as AIC or BIC is well explored and understood for models that correspond to fixed-effects likelihoods. However, in the presence of random effects their use can be problematic. For model selection in time series models, the comparison of successive one-step-ahead forecasts with the actually observed data provides a natural alternative. In this context, Gneiting and Raftery (2007) recommend the use of strictly proper scoring rules, such as the logarithmic score or the ranked probability score. See Czado et al. (2009) and Paul and Held (2010) for further details.

One-step-ahead predictions for the last 2 years for model B2 are obtained as follows:

```
> pred.B2 <- oneStepAhead(result.B2, tp = nrow(flu) - 2 * 52)
```

The mean logarithmic and mean ranked probability score are then computed with

```
> colMeans(scores(pred.B2)[, c("logs", "rps")])
```

```
     logs       rps
0.5632647 0.4362529
```

As a last example, consider the number of measles cases in the 16 federal states of Germany, in the years 2005–2007. There is considerable regional variation in the incidence pattern which is most likely due to differences in vaccination coverage. In the following, information about vaccination coverage in each state, namely the log proportion of unvaccinated school starters, is included as explanatory variable in a model for the bi-weekly aggregated measles data. See Herzog et al. (2010) for further details.
The 78×16 matrix `vac0` contains the proportion of unvaccinated school starters in each state i.

```
> vac0[1:2, 1:5]
```

```
     Baden-Württemberg  Bavaria   Berlin Brandenburg   Bremen
[1,]         0.1000115 0.113261 0.099989   0.0605575 0.115963
[2,]         0.1000115 0.113261 0.099989   0.0605575 0.115963
```

A Poisson model which links the autoregressive parameter with this covariate and contains $S = 1$ seasonal term in the endemic component (cf. model A0 in Tab. 3 in Herzog et al., 2010) is obtained with

```
> f.end <- addSeason2formula(f = ~ 1, S = 1, period = 26)
> # autoregressive component: Intercept + vaccination coverage information
> model.A0 <- list(ar = list(f = ~ 1 + logVac0),
+                  end = list(f = f.end, offset = population(measles2w)),
+                  data = list(t = epoch(measles2w), logVac0 = log(vac0)))
> # fit model
> result.A0 <- hhh4(measles2w, model.A0)
> # parameter estimates
> round(coef(result.A0,
+            se = TRUE,              # also return standard errors
+            amplitudeShift = TRUE   # transform sin/cos terms to
+            ), 2)                   # Amplitude/shift formulation
```

```
                     Estimates Std. Error
ar.1                      3.01       0.52
ar.logVac0                1.38       0.23
end.1                     1.78       0.06
end.A(2 * pi * t/26)      0.66       0.08
end.s(2 * pi * t/26)     -0.10       0.12
```

5 Summary

As part of the R-package `surveillance`, the function `hhh4` provides a flexible tool for the modelling of multivariate time series of infectious disease counts. The discussed count data model is able to account for serial and spatio-temporal correlation, as well as heterogeneity in incidence levels and disease transmission. The functionality of `hhh4` was illustrated using several built-in data sets.

References

Czado, C., Gneiting, T., and Held, L. (2009). Predictive model assessment for count data. *Biometrics*, 65(4):1254–1261.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Held, L., Höhle, M., and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling*, 5(3):187–199.

Herzog, S. A., Paul, M., and Held, L. (2010). Heterogeneity in vaccination coverage explains the size and occurrence of measles epidemics in German surveillance data. *Epidemiology and Infection*. doi:10.1017/S0950268810001664.

Höhle, M. (2007). Surveillance: an R package for the monitoring of infectious diseases. *Computational Statistics*, 22(4):571–582.

Höhle, M. and Mazick, A. (2010). Aberration detection in R illustrated by Danish mortality monitoring. In Kass-Hout, T. and Zhang, X., editors, *Biosurveillance: A Health Protection Priority*. CRC Press.

Höhle, M., Riebler, A., and Paul, M. (2007). The R-package 'surveillance'.

Lewin-Koh, N. J., Bivand, R., contributions by Edzer J. Pebesma, Archer, E., Baddeley, A., Bibiko, H.-J., Dray, S., Forrest, D., Friendly, M., Giraudoux, P., Golicher, D., Rubio, V. G., Hausmann, P., Jagger, T., Luque, S. P., MacQueen, D., Niccolai, A., Short, T., and Stabler, B. (2010). *maptools: Tools for reading and handling spatial objects*. R package version 0.7-34.

Paul, M. and Held, L. (2010). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Statistics in Medicine*. Accepted.

Paul, M., Held, L., and Toschke, A. M. (2008). Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine*, 27(29):6250–6267.

Robert Koch-Institut (2009). SurvStat@RKI. `http://www3.rki.de/SurvStat`. Accessed March 2009.